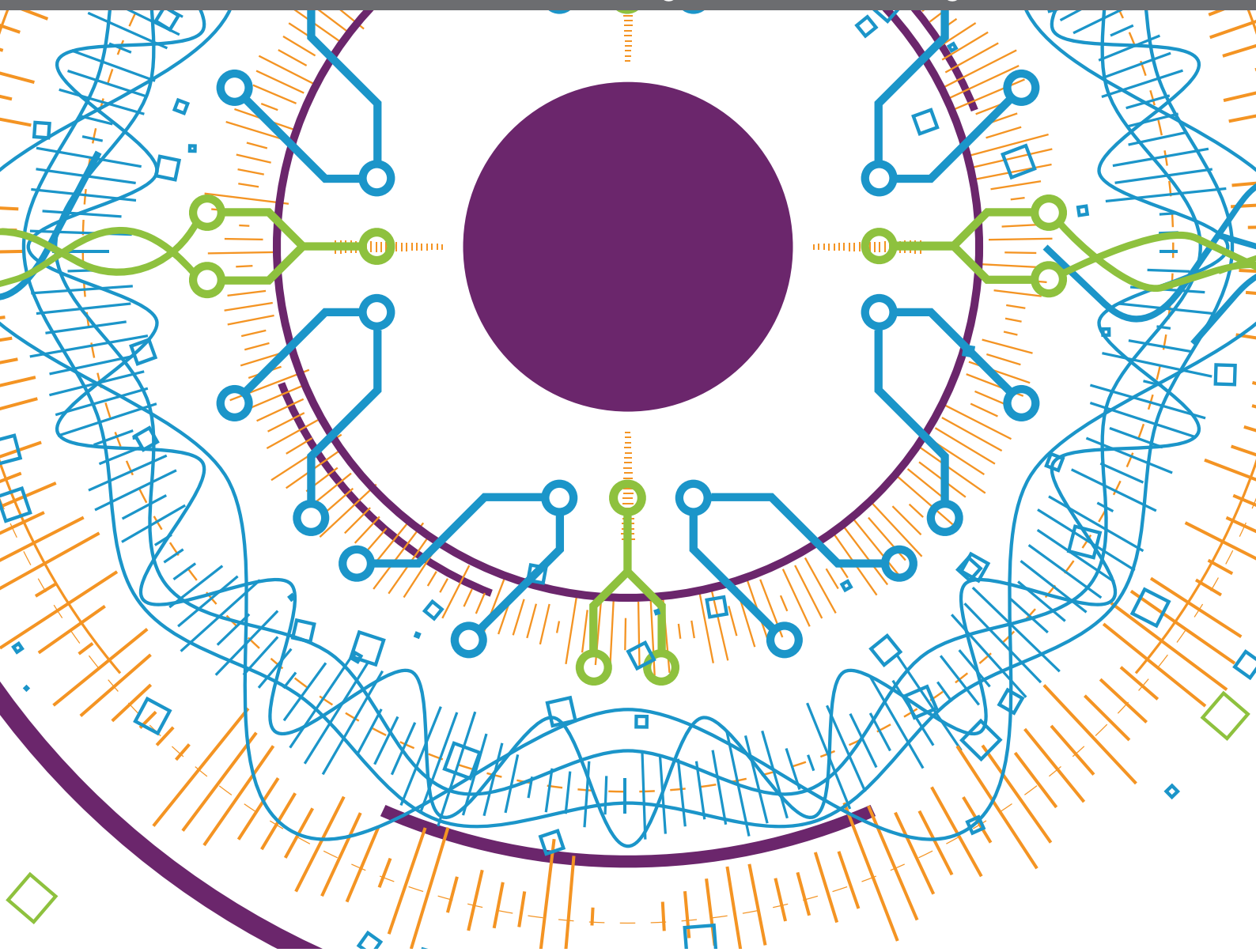
PROBABILISTIC PERSPECTIVES ON BRAIN (dys)FUNCTION

EDITED BY: Karl Friston, Thomas Parr, Dimitrije Marković,
Maxwell James D. Ramstead, Ryan Smith and Casper Hesp

PUBLISHED IN: Frontiers in Artificial Intelligence and Frontiers in Big Data



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-128-4
DOI 10.3389/978-2-88971-128-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

PROBABILISTIC PERSPECTIVES ON BRAIN (dys)FUNCTION

Topic Editors:
Karl Friston, University College London, United Kingdom
Thomas Parr, University College London, United Kingdom
Dimitrije Marković, Technische Universität Dresden, Germany
Maxwell James D. Ramstead, McGill University, Canada
Ryan Smith, Laureate Institute for Brain Research, United States
Casper Hesp, Amsterdam Brain and Cognition (ABC), Netherlands

Citation: Friston, K., Parr, T., Marković, D., Ramstead, M. J. D., Smith, R., Hesp, C., eds. (2021). Probabilistic Perspectives on Brain (dys)function. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-128-4

Table of Contents

frontiers in Artificial Intelligence

EDITORIAL
published: 07 June 2021
doi: 10.3389/frai.2021.710179

Editorial: Probabilistic Perspectives on Brain (Dys)function

Thomas Parr[1]*, *Dimitrije Marković*[2], *Maxwell James D. Ramstead*[1,3,4,5], *Ryan Smith*[6], *Casper Hesp*[1,7] and *Karl Friston*[1]

[1]Wellcome Centre for Human Neuroimaging, Queen Square Institute of Neurology, University College London, London, United Kingdom, [2]Department of Psychology, Technische Universität Dresden, Dresden, Germany, [3]Division of Social and Transcultural Psychiatry, Department of Psychiatry, McGill University, Montreal, QC, Canada, [4]Spatial Web Foundation, Los Angeles, CA, United States, [5]Nested Minds Network, London, United Kingdom, [6]Laureate Institute for Brain Research, Tulsa, OK, United States, [7]Amsterdam Brain and Cognition Center, University of Amsterdam, Amsterdam, Netherlands

Keywords: neuroscience, artificial intelligence, computational psychiatry, Bayesian inference, generative models

Editorial on the Research Topic

Probabilistic Perspectives on Brain (Dys)function

While observations in neurobiology provide inspiration for methods in artificial intelligence and machine learning—most famously, in the development of artificial neural networks (McCulloch and Pitts 1943; Rosenblatt 1958; Smolensky 1986) —the reciprocal relationship has also proved fruitful. Put simply, many of the problems that machine learning is designed to solve have already been solved by the brain. When we have a good understanding of how the brain deals with a problem, we can draw inspiration from this solution in other domains. When we have a poor understanding of aspects of brain function, we can look to how these functions are performed in machine learning. If natural selection has arrived at the same optimum, we hypothesize that brain architectures support analogous procedures. Perhaps the most obvious example of this translation is the Bayesian brain hypothesis (Knill and Pouget 2004; Doya 2007), and recent extensions of this idea (Ramstead et al., 2018). This perspective treats the brain as a statistician who makes use of a probabilistic model of the world to make sense of sensory input. It has been central to the development of theories of brain function—like predictive coding (Srinivasan et al., 1982; Rao and Ballard 1999; Friston and Kiebel 2009; Bastos et al., 2012). This research topic was designed to showcase the application of contemporary probabilistic methods to understanding how the brain works, and how it can go awry in psychiatric disorders.

Broadly, the applications of probabilistic methods to the brain fall into two camps. The first applies these methods to neurobiological or psychophysical data to draw better inferences about the brain. The second assumes the brain itself makes use of these methods and engages in inference about the data it gathers from receptors in the eyes, ears, and other sensory organs. Both approaches are usefully illustrated by Feltgen and Daunizeau. Their focus is on refinement of the estimation procedure for drift-diffusion models (Ratcliff and McKoon, 2008). While drift-diffusion dynamics may be seen as a metaphor for evidence accumulation in the brain, the estimation procedure advocated by the authors represents a means of drawing inferences about cognition from psychophysical measurements.

A related perspective on evidence accumulation is offered by Heins et al., who show the emergence of drift-diffusion like dynamics in belief updating under a deep temporal model (Friston et al., 2017). This introduces an active aspect, in which we must decide how to sample our sensory data, over multiple timescales, to ensure we assimilate the most informative data (Mirza et al., 2016). The neural realization of this assimilation process was probed by Loued-Khenissi and Preuschoff in a functional imaging experiment in which participants engaged in a probabilistic gambling task. The task allowed the authors to disambiguate neural correlates of the confidence with which an outcome was predicted from the information gain when it is observed.

OPEN ACCESS

Edited and reviewed by:
Thomas Hartung,
Johns Hopkins University,
United States

***Correspondence:**
Thomas Parr
thomas.parr.12@ucl.ac.uk

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *15 May 2021*
Accepted: *24 May 2021*
Published: *07 June 2021*

Citation:
Parr T, Marković D, Ramstead MJD, Smith R, Hesp C and Friston K (2021). Editorial: Probabilistic Perspectives on Brain (Dys)function Front. Artif. Intell. 4:710179. doi: 10.3389/frai.2021.710179

Chen et al. exploit the same active inferential formalism as Heins et al., but apply it to understand how the brain might optimize the space of hypotheses it entertains. Specifically, the authors employ Bayesian model reduction (Friston et al., 2016; Friston et al., 2018)—a technique originally developed to compare dynamic causal models in neuroimaging—to prune the set of behavioral policies a creature can select between. Policies here are alternative sequences (of actions) over time. These could be sequences of saccadic eye movements, or steps through a maze (Kaplan and Friston, 2018). Such sequences are ubiquitous in planning and decision-making problems.

Temporal sequences of this sort are central to two other contributions to this Research Topic. Frölich et al. review the generation of sequences in neural systems in the form of robust and reproducible activation patterns and argue for their central role in probabilistic and predictive information processing. FitzGerald et al. complement this by considering the role of retrospective (postdictive) inference; through the perspective of Bayesian filtering (prospective) and smoothing (prospective and retrospective). The authors propose a middle ground between the two by limiting the number of past time-steps over which retrospective inference is performed—curtailing the computational cost accrued in modeling long sequences—and demonstrate the success of the resulting scheme on a probabilistic reversal learning task.

At a more conceptual level, Safron provides a broad overview of active inference and its relationship to other influential theories of brain and consciousness, including the global neuronal workspace theory (Baars, 1993) and integrated information theory (Tononi et al., 2016). Gershman adds an interesting novel perspective to this through proposing a generative adversarial theory of brain function. This is based upon the widely used deep learning networks of the same name (Goodfellow et al., 2014). Generative adversarial networks learn a generative model of the data they are exposed to. Their objective is to generate new data that are indistinguishable from the original inputs. Gershman highlights how human brain architectures could support the generative and discriminative parts of such networks.

A key area of application for theoretical neurobiology is in computational psychiatry (Montague et al., 2012). This interdisciplinary field is well-represented by the contributions from Leptourgos and Corlett and Mehltretter et al. The former set out a theory for the distortions in the sense of agency experienced by some people with schizophrenia. They do so through assuming the brain makes use of two distinct predictive hierarchies that deal with the feeling of, and the judgment of, agency, respectively. This dual hierarchy allows them to incorporate features of prominent theories of passivity phenomena (Blakemore and Frith 2003; Synofzik et al., 2008). Mehltretter et al. take a different perspective on computational psychiatry and make use of deep learning methods in feature selection to predict remission of symptoms in patients taking antidepressants. Their focus is on the important challenge of interpretability for such analyses.

The papers outlined above offer a snapshot of the exciting work at the interface of neuroscience and probabilistic reasoning and the enduring symbiotic relationship between the two fields.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

DM was funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft), SFB 940/2, 543 project A9. KF was a Wellcome Principal Research Fellow (Ref: 088130/Z/09/Z). RS is supported by the William K. Warren Foundation, the Stewart G. Wolf Fellowship, and a Center Grant from the National Institute of General Medical Sciences (P20GM121312). Postdoctoral Fellowship from the Social Sciences and Humanities Research Council of Canada (Ref: 756-2020-0704) (MR).

ACKNOWLEDGMENTS

We are grateful to the authors who contributed their work to this special issue, and to the peer reviewers for their invaluable assistance in evaluating the submissions.

REFERENCES

Baars, B. J. (1993). *A Cognitive Theory of Consciousness*. Cambridge, United Kingdom: Cambridge University Press.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron* 76 (4), 695–711. doi:10.1016/j.neuron.2012.10.038

Blakemore, S.-J., and Frith, C. (2003). Self-awareness and Action. *Curr. Opin. Neurobiol.* 13 (2), 219–224. doi:10.1016/s0959-4388(03)00043-6

Doya, K. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Cambridge, MA: MIT press.

Friston, K. J., Litvak, V., Oswal, A., Razi, A., Stephan, K. E., van Wijk, B. C. M., et al. (2016). Bayesian Model Reduction and Empirical Bayes for Group (DCM) Studies. *NeuroImage* 128 (Suppl. C), 413–431. doi:10.1016/j.neuroimage.2015.11.015

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017). Deep Temporal Models and Active Inference. *Neurosci. Biobehavioral Rev.* 77, 388–402. doi:10.1016/j.neubiorev.2017.04.009

Friston, K., and Kiebel, S. (2009). Predictive Coding under the Free-Energy Principle. *Phil. Trans. R. Soc. B* 364 (1521), 1211–1221. doi:10.1098/rstb.2008.0300

Friston, K., Parr, T., and Zeidman, P. (2018). "Bayesian Model Reduction." arXiv preprint arXiv:1805.07092.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative Adversarial Networks." arXiv preprint arXiv:1406.2661.

Kaplan, R., and Friston, K. J. (2018). *Planning and Navigation as Active Inference*. Biological Cybernetics.

Knill, D. C., and Pouget, A. (2004). The Bayesian Brain: the Role of Uncertainty in Neural Coding and Computation. *Trends. Neurosci.* 27 (12), 712–719. doi:10.1016/j.tins.2004.10.007

McCulloch, W. S., and Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull. Math. Biophys.* 5 (4), 115–133. doi:10.1007/bf02478259

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016). Scene Construction, Visual Foraging, and Active Inference. *Front. Comput. Neurosci.* 10 (56), 1–16. doi:10.3389/fncom.2016.00056

Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational Psychiatry. *Trends Cogn. Sci.* 16 (1), 72–80. doi:10.1016/j.tics.2011.11.018

Ramstead, M. J. D., Badcock, P. B., and Friston, K. J. (2018). Variational Neuroethology: Answering Further Questions. *Phys. Life Rev.* 24, 59–66. doi:10.1016/j.plrev.2018.01.003

Rao, R. P. N., and Ballard, D. H. (1999). Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-Field Effects. *Nat. Neurosci.* 2 (1), 79–87. doi:10.1038/4580

Ratcliff, R., and McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Comput.* 20 (4), 873–922. doi:10.1162/neco.2008.12-06-420

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychol. Rev.* 65 (6), 386–408. doi:10.1037/h0042519

Smolensky, P. (1986). "Information Processing in Dynamical Systems: Foundations of harmony Theory," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (foundations: MIT Press), Vol. 1, 194–281.

Srinivasan, M. V., Laughlin, S. B., Dubs, A., and Horridge, G. A. (1982). Predictive Coding: a Fresh View of Inhibition in the Retina. *Proc. R. Soc. Lond. B. Biol. Sci.* 216 (1205), 427–459. doi:10.1098/rspb.1982.0085

Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the Comparator Model: A Multifactorial Two-step Account of agency. *Conscious. Cogn.* 17 (1), 219–239. doi:10.1016/j.concog.2007.03.010

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated Information Theory: from Consciousness to its Physical Substrate. *Nat. Rev. Neurosci.* 17 (7), 450–461. doi:10.1038/nrn.2016.44

Conflict of Interest: MR was employed by the company Spatial Web Foundation and Nested Minds Network.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

HYPOTHESIS AND THEORY
published: 18 September 2019
doi: 10.3389/frai.2019.00018

The Generative Adversarial Brain

*Samuel J. Gershman**

Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA, United States

The idea that the brain learns generative models of the world has been widely promulgated. Most approaches have assumed that the brain learns an explicit density model that assigns a probability to each possible state of the world. However, explicit density models are difficult to learn, requiring approximate inference techniques that may find poor solutions. An alternative approach is to learn an implicit density model that can sample from the generative model without evaluating the probabilities of those samples. The implicit model can be trained to fool a discriminator into believing that the samples are real. This is the idea behind generative adversarial algorithms, which have proven adept at learning realistic generative models. This paper develops an adversarial framework for probabilistic computation in the brain. It first considers how generative adversarial algorithms overcome some of the problems that vex prior theories based on explicit density models. It then discusses the psychological and neural evidence for this framework, as well as how the breakdown of the generator and discriminator could lead to delusions observed in some mental disorders.

Keywords: bayesian inference, delusions, consciousness, generative adversarial networks, perception

OPEN ACCESS

Edited by:
Thomas Parr,
University College London,
United Kingdom

Reviewed by:
Alexander Daniel Dunsmoir Tschantz,
University of Sussex, United Kingdom
Bobbie-Jo Webb-Robertson,
Pacific Northwest National Laboratory
(Department of Energy), United States

***Correspondence:**
Samuel J. Gershman
gershman@fas.harvard.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *22 July 2019*
Accepted: *02 September 2019*
Published: *18 September 2019*

Citation:
Gershman SJ (2019) The Generative Adversarial Brain. Front. Artif. Intell. 2:18. doi: 10.3389/frai.2019.00018

1. INTRODUCTION

Our sensory inputs are impoverished, and yet our experience of the world feels richly detailed. For example, our fovea permits us access to a high fidelity region of the visual field only twice the size of our thumbnail held at arm's length. But we don't experience the world as though looking through a tiny aperture. Instead, our brains feed us a "grand illusion" of panoptic vision (Noë et al., 2000; Chater, 2018; Odegaard et al., 2018). Similarly, we receive no visual input in the region of the retina that connects to the optic nerve, yet under normal circumstances we are unaware of this blind spot. Moreover, even when we receive high fidelity visual input, we may still fail to witness dramatic changes in scenes (Simons, 2000), as though our brains have contrived imaginary scenes that displace the true scenes.

There is a standard inferential explanation of these and many other illusions (e.g., Gregory, 1980), which holds that our percepts reflect beliefs about the world rather than raw sensory information. In modern computational models of perception, these beliefs are typically conceptualized as probability distributions over some hypothesis space conditional on the sensory input, as stipulated by Bayes' rule (Knill and Richards, 1996):

$$P(z|x) = \frac{P(x|z)P(z)}{\sum_{z'} P(x|z')P(z')}, \tag{1}$$

where $P(x|z)$ is the likelihood of the data x given hypothesis z, $P(z)$ is the prior probability of z, and $P(z|x)$ is the posterior probability. While the Bayesian framework has considerable merit, it does not seem to provide adequate answers to several questions.

First, how can we explain the phenomenology of illusion: why do some illusions feel *real*, as though one is actually seeing them, whereas other inferences carry information content without

the same perceptual experience. For example, Ramachandran and Hirstein (1997) use the example of gazing at wallpaper in a bathroom, where the wallpaper in your visual periphery is "filled in" (you subjectively experience it as high fidelity even though objectively you perceive it with low fidelity), but the wallpaper behind your head is not filled in. In other words, you *infer* that the wallpaper continues behind your head, and you may even know this with high confidence, but you do not have the experience of *seeing* the wallpaper behind your head. Thus, the vividness or "realness" of perceptual experience is not a simple function of belief strength. So what is it a function of?

Second, how can we explain the peculiar ways that the inferential apparatus breaks down? In particular, how can we understand the origins of delusions, hallucinations, and confabulations that arise in certain mental disorders? While Bayesian models have been developed to explain these phenomena, they fall short in certain ways that we discuss later on.

In this paper, we argue that these issues can be addressed by thinking about Bayesian inference from a different algorithmic perspective. The basic idea is that a "generator" draws samples from the generative model, which are then fed, along with samples of real sensory data, into a "discriminator" that tries to figure out which samples are real and which are fake. These two components are in a kind of arms race: the generator is trying to produce samples that trick the discriminator into incorrectly classifying them as real, and the discriminator is trying to learn how to detect these fakes. If the visual system plays the role of the generator, and our perceptual experience reflects the judgment of the discriminator, then we can begin to understand why the visual system might report things that aren't there, or fail to report things that are there, and why our perceptual experience endorses these false or incomplete reports (see also Lau, 2019). Furthermore, breakdown of the generator and discriminator may explain the origin of false beliefs and percepts in certain mental disorders: a dysfunctional generator can produce abnormal content, and a dysfunctional discriminator can endorse that content as real.

This "generative-adversarial" interplay is motivated by recent advances in machine learning, which have produced algorithms for learning generative models based on the same idea. In the next section, we summarize the idea more formally. What follows is a rampantly speculative discussion of implications for psychology and neuroscience (note that the article is not proposing any novel computational ideas from the perspective of machine learning). Finally, we apply these ideas to understanding delusions observed in some mental disorders.

2. GENERATIVE MODELS: EXPLICIT AND IMPLICIT

Generative models can be understood as stochastic "recipes" for generating observed data: first draw a latent variable z from the prior $P(z)$, then draw data from the conditional distribution $P(x|z)$. This generative model can then be inverted according to Bayes' rule to recover a posterior belief $P(z|x)$ about the latent variable conditional on the data. There are two basic problems that any probabilistic information processing system (artificial or biological) must face. The *inference problem* is how to compute the posterior efficiently given constraints on computational resources. The *learning problem* is to update the generative model $P(x, z)$ in order to better match the empirical data distribution. Learning is limited both by the amount of training data and by the difficulty of searching through the space of probability distributions (typically via gradient-based techniques).

Exact Bayesian inference is intractable for most moderately complex generative models. This means that if we are going to consider expressive generative models, we will need to also consider approximate inference. Historically, approximate inference algorithms have fallen into two families (Gershman and Beck, 2017). One family, Monte Carlo algorithms, approximates the posterior via stochastic simulation. Provided enough samples are drawn, Monte Carlo algorithms can, at least in theory, approximate the posterior arbitrarily well. They can account for a wide range of neural (Buesing et al., 2011; Haefner et al., 2016; Orbán et al., 2016), and behavioral (Sanborn and Chater, 2016; Dasgupta et al., 2017) data. Their main limitation is that they can be woefully inefficient for complex distributions, unless one uses more sophisticated variants that pose challenges for neural and psychological plausibility.

The second family, variational algorithms, approximate the posterior with a simpler parameterized form that is easier to optimize. Variational algorithms have figured prominently in neuroscience, where they underpin the free-energy principle (Friston, 2009), and have also been proposed as psychologically plausible process models (Sanborn and Silva, 2013; Dasgupta et al., 2019). These algorithms are often much more efficient compared to Monte Carlo, which is why they are widely used in machine learning. However, because of the simplified parameterization, the optimal approximation will typically be biased (i.e., it won't perfectly capture the true posterior).

A basic limitation of both Monte Carlo and variational algorithms is that they are mainly designed to work with *explicit* generative models: they assume that the likelihood can be evaluated for any data sample. However, there are many complex models that are *implicit* in the sense that they can only be simulated. For example, the drift-diffusion model does not have a tractable closed-form expression for the likelihood function, but samples can be drawn from the generative model. This has motivated various forms of "likelihood-free" algorithms (e.g., Diggle and Gratton, 1984; Csilléry et al., 2010; Hartig et al., 2011; Gutmann and Corander, 2016).

Recently, a new approach to likelihood-free approximate inference has emerged based on a minimax game between a generator G and a discriminator D (Donahue et al., 2016; Dumoulin et al., 2017).[1] Both the generator and discriminator are typically implemented as differentiable neural networks.

[1]The space of generative-adversarial algorithms is much broader than what is covered in this paper. The original formulation (which did not involve inference at all) is due to Goodfellow et al. (2014). The relationship between generative-adversarial inference algorithms and other approximate inference algorithms is discussed in Huszár (2017).

The discriminator takes as input data x and latent variable z, and outputs the probability that (x, z) was drawn from the joint distribution $P(x, z)$ vs. the generator distribution $G(x, z)$. The generator consists of two components (**Figure 1**): a "feedforward" component $G(z|x)$ that samples inferred latent variables $\hat{z}$ conditional on empirical data $x \sim P(x)$, and a "feedback" component $G(x|z)$ that samples simulated data $\hat{x}$ conditional on draws from the prior $z \sim P(z)$. The feedforward component implements the approximate inference engine, efficiently mapping data to samples from the approximate posterior over latent variables. The feedback component implements the learned generative model, mapping latent variables to samples from the observation distribution.

The generator and discriminator are jointly trained to optimize the following "adversarial" objective function:

$$\min_G \max_D \mathbb{E}_{G(z|x)P(x)} \left[\log D(x, z)\right] + \mathbb{E}_{G(x|z)P(z)} \left[\log(1 - D(x, z))\right]. \tag{2}$$

Intuitively, the generator is trying to fool the discriminator into placing high probability on simulated data and low probability on empirical data, while the discriminator is trying to do the opposite. It can be shown (Dumoulin et al., 2017) that the optimal discriminator for a fixed generator is given by:

$$D^*(x, z) = \frac{G(x, z)}{G(x, z) + P(x, z)}. \tag{3}$$

Thus, the discriminator will be at chance when the generator has perfectly approximated the true joint distribution. The optimal generator can also be understood as minimizing the Jensen-Shannon divergence between G and P (Goodfellow et al., 2014; Dumoulin et al., 2017)[2].

Adversarially learned inference has two important advantages over standard Monte Carlo and variational approaches. First, as already noted, it can be applied to implicit generative models, which means that these models can be more complex (e.g., parameterized as a deep neural network with an intractable likelihood function). The result is that the quality of the generative model is higher, as measured (for example) in terms of simulated data quality. Second, inference is more efficient than standard Monte Carlo algorithms (it is "amortized" in the form of a learned function that can be quickly evaluated) and can use more flexible approximate posteriors compared to standard variational algorithms[3].

3. PSYCHOLOGICAL IMPLICATIONS

3.1. The Puzzle of Phenomenology

We began this paper with examples from visual perception in which people have the subjective experience of seeing things that are objectively not there (e.g., high acuity in the periphery or in the retinal blind spot). This is sometimes discussed as perceptual "filling-in," though this term is theoretically tendentious: it suggests something like a neural paintbrush that fills in missing segments on an internal screen, an idea that (Dennett, 1992) has argued is highly implausible. As an alternative, Dennett suggests something more like "paint-by-numbers," where surfaces are symbolically labeled, and these symbols are interpreted appropriately by downstream computations. Indeed, this is roughly how digital computers typically deal with surfaces.

As a matter of neurophysiology, it turns out that Dennett was incorrect: there really is an interpolation process in low-order visual areas that is retinotopically organized (De Weerd, 2006). The more important point for present purposes is that Dennett's argument doesn't really explain the subjective experience of perceptual filling-in. Either interpolative or symbolic implementations could be compatible with this subjective experience. In essence, the question is why the downstream interpreter of these representations ascribes "realness" to some representations (wallpaper in front of you, to again use Ramachandran and Hirstein's example) and not others (wallpaper behind you).

Noë et al. (2000) have offered a different line of argument, that we don't actually have the subjective experience of seeing stimuli in the periphery or the blind spot, but rather our phenomenology reflects the knowledge that the relevant stimulus information is available in the environment, and we could (e.g., with eye movements) apprehend that information. This seems somewhat unsatisfactory, because it is basically denying the introspective observation that we experience ourselves as really seeing stimuli in the periphery. It also seems to conflict with psychophysical experiments demonstrating that people are overconfident about how much they see in the periphery (Odegaard et al., 2018). If it was simply a matter of knowing that we *could* see something, not that we actually *do* see something, then there's no reason why we should feel overconfident about our perceptual acuity.

The adversarial framework leads to another way of thinking about these issues. The discriminator is, by design, making ascriptions of "realness" to inputs that are both real and simulated. Meanwhile, the generator is trying its best to feed the discriminator realistic simulations. Thus, if subjective perceptual experience corresponds to perceptual content that has been endorsed as real by the discriminator, then we would have an explanation for why we feel that we see more than we do. Simulations of peripheral visual input are highly compelling. On the other hand, simulations of visual inputs outside the field of vision are not. The generator can trick the discriminator into thinking that it sees wallpaper in front of us, but not behind us.

This perspective has some resonance with higher-order theories of consciousness (Lau and Rosenthal, 2011; Lau, 2019), which hold that conscious awareness is a particular kind of mental state that represents other mental states. The discriminator can be understood as a higher-order representation that represents beliefs (real vs. imagined) about lower-level perceptual representations. On this view, conscious awareness

[2]Note that the product rule of probability implies that $G(x, z) = G(z|x)P(x) = G(x|z)P(z)$. However, because the two generator components are parameterized independently, this equality may not hold in practice, except at the optimum of the objective function (provided both components are sufficiently expressive).

[3]Note that amortization can also be applied to variational inference in explicit generative models, so this advantage is not unique.

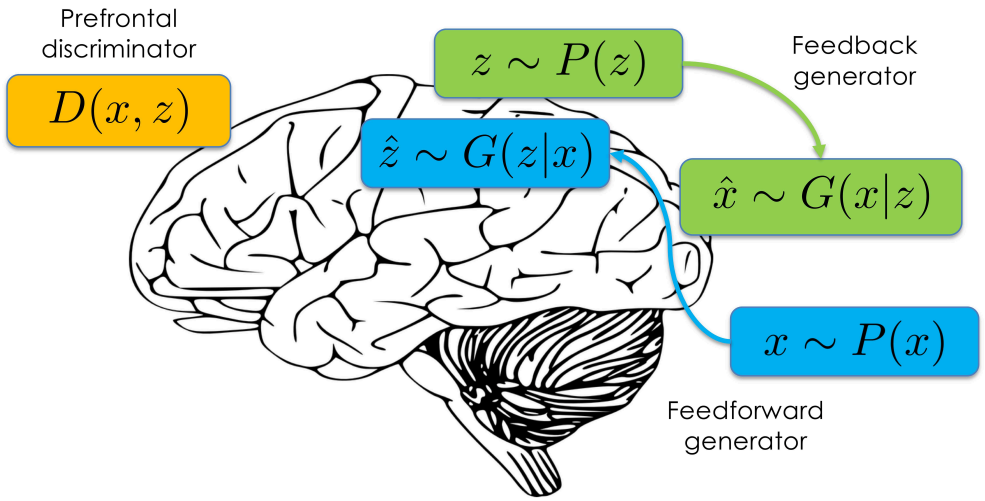


FIGURE 1 | Schematic of the adversarially learned inference architecture mapped onto the brain.

occurs when a decision is made that a perceptual representation is veridical (see also Dehaene et al., 2014).

The adversarial framework contrasts with the interoceptive predictive coding account of Seth et al. (2012), according to which the sense of reality derives from the perception of sensorimotor contingency. While sensorimotor contingency might be one piece of information that the discriminator uses to make its decisions, it can also use other sources of information. For example, people who are unable to move their eyes may experience low sensorimotor contingency, but can still discriminate real from imagined stimuli.

3.2. Discriminating Between Reality and Imagination

The adversarial framework posits that a mechanism for discriminating between reality and imagination plays an important computational role in learning and inference. In the psychology literature, the discrimination problem has been studied in the context of *reality testing* (discriminating between real and imagined stimuli in perception) and *reality monitoring* (discriminating between real and imagined stimuli in memory). The most famous example of reality testing is the Perky effect. Perky (1910) presented subjects with dimly illuminated images of objects while subjects were asked to describe the objects, and found that subjects falsely reported these as imagery rather than perception. Segal and Fusella (1970) examined this effect with signal detection techniques, finding that sensitivity was reduced under mental imagery conditions, particularly for perceived and imagined stimuli in the same sensory modality. Many subsequent studies have documented interactions between imagery and perception. For example, Farah and Smith (1983) demonstrated that imagery can facilitate stimulus detection (see also Farah, 1985; Ishai and Sagi, 1995).

The study of reality monitoring has been championed by Johnson and her collaborators (see Johnson and Raye, 1981, for a review of the early literature), who have called attention to the problem that mental images leave traces in memory, and therefore some mechanism must exist to discriminate between these memories and memories of observed stimuli. As we discuss below, this mechanism appears to have a dedicated neural substrate, and dysfunction of this mechanism may underpin cognitive and perceptual symptoms in certain mental disorders. One important set of findings from research on reality monitoring is the identification of factors that people use to discriminate reality from imagination. For example, real stimuli are richer in perceptual and semantic detail, and contain less information about cognitive operations. These are all factors we would expect that a well-designed discriminator could exploit.

4. NEURAL IMPLICATIONS

The architecture shown in **Figure 1** lends itself naturally to a systems-level interpretation. The discriminator corresponds to a reality monitoring mechanism that has been frequently attributed to the median anterior prefrontal cortex (see Simons et al., 2017, for a review). For example, this region is activated when subjects are asked to discriminate whether a visual object was previously seen or imagined (Kensinger and Schacter, 2006), and morphological features of this region covary with individual differences in reality monitoring performance (Buda et al., 2011). Moreover, patients with schizophrenia (Garrison et al., 2017) and healthy individuals prone to expression of psychotic and schizotypal traits (Simons et al., 2008) both show reduced activation in this area during reality monitoring.

The "feedback" and "feedforward" terminology was chosen to suggest a mapping onto feedback and feedforward pathways in posterior cortical regions. This is consistent with theories of cortical function that posit a role for feedforward pathways in computing inferences about the latent causes of sensory data, and a role for feedback pathways in computing predictions about upcoming sensory data (e.g., Dayan et al., 1995; Lee and Mumford, 2003; Lochmann and Deneve, 2011). Some

theories (e.g., Rao and Ballard, 1999; Friston, 2008) have argued that feedforward pathways convey prediction *errors* rather than predictions. This can be understood as an efficient way to pass predictions up the cortical hierarchy while removing redundant information (see Huang and Rao, 2011).

At the circuit level, an implicit generative model could be implemented as a probabilistic population code (PPC; Ma et al., 2006), which represents a probability distribution via the distribution of spikes across a population. One challenge facing PPCs is that they only support exact inference for relatively simple generative models, such as Kalman filtering and multi-sensory cue combination. Some authors have attempted to generalize PPCs to the approximate inference setting, for example by having the PPCs encode the sufficient statistics of a factorized variational approximation (Beck et al., 2012) or the sufficient statistics of cliques in a graphical model that then pass messages using loopy belief propagation (Raju and Pitkow, 2016). Both of these generalizations limit the kinds of generative models that can be represented. Adversarially learned inference provides potentially another way to work with more flexibly parameterized models. An open problem is to determine what kinds of biologically plausible learning rules could implement optimization of the adversarial objective function.

5. DELUSIONS

In the field of cognitive neuropsychiatry, some authors have invoked inferential explanations of delusion formation (Hemsley and Garety, 1986; Corlett et al., 2009; Coltheart et al., 2010; McKay, 2012; Sterzer et al., 2018). According to the "two-factor" version of this idea (see Coltheart et al., 2010), two underlying factors must break down: (i) the input data must be abnormal, and (ii) the hypotheses suggested by the abnormal data must be defectively evaluated. Some patients have an impaired first factor but an intact second factor; these patients have abnormal experiences but do not develop delusions. Coltheart et al. (2010) viewed the evaluation factor as a form of Bayesian inference, but conceded that Bayes' rule is silent about the origin of abnormal data (the first factor). Moreover, the conjectured impairment in the evaluation factor—that patients are unable to assimilate evidence contradicting the delusional belief—runs into trouble. As pointed out by McKay (2012), it doesn't really make sense chronologically why patients would be able to assimilate the abnormal data but not the subsequent contradictory data. As an alternative, McKay suggests that the impairment in the evaluation factor is a bias toward "explanatory adequacy," whereby the likelihood is overweighted at the expense of the prior. This alternative still leaves the origin of abnormal data unexplained.

In support of the two-factor interpretation, Coltheart et al. (2010) discuss evidence that impairments of abnormal data and abnormal evaluation are dissociable. For example, some patients with damage to the ventromedial prefrontal cortex fail to autonomically discriminate between familiar and unfamiliar faces, as measured by skin conductance, despite their ability to recognize the familiar faces (Tranel et al., 1995). Coltheart et al. view these cases as analogous to Capgras patients, in the sense that both syndromes produce abnormal content, but with the critical difference that Capgras patients develop delusions because of their impaired ability to evaluate the abnormal content, whereas ventromedial prefrontal patients do not develop delusions.

Another example is the Fregoli delusion, which is essentially the opposite of the Capgras delusion: patients perceive strangers as familiar people in disguise. It has been suggested that the underlying mechanism of abnormal content generation is the opposite of the putative mechanism underlying Capgras delusion, namely an over-responsive autonomic response to faces (Ramachandran et al., 1998). Importantly, there are patients who show the same abnormal content generation (strange faces are perceived as highly familiar) but who do not develop delusions (Vuilleumier et al., 2003).

Some theorists have advocated for a "one-factor" predictive coding version of the inferential account (e.g., Corlett et al., 2009; Sterzer et al., 2018), according to which delusion formation arises from a single cause: noisy prediction errors, which register the discrepancy between observations and expectations and drive updating of beliefs. Noise in the prediction errors furnishes the abnormal input data, which in turn drives aberrant belief updating. One potentially problematic aspect of this account is that it seems to require the noise to be quite large in order to produce the kinds of dramatic delusions that have been observed (e.g., believing that family members have been replaced by imposters, as in Capgras syndrome). Although there is evidence for noisy neural signaling in schizophrenia (Winterer and Weinberger, 2004), signal detection analyses of psychophysical performance have indicated that internal noise levels do not differ between schizophrenics and healthy controls (Collicutt and Hemsley, 1981; Bentall and Slade, 1985). Moreover, some disorders (e.g., autism; see Dinstein et al., 2012; Park et al., 2017) have been associated with elevated noise levels but are not reliably associated with delusions (though see van Schalkwyk et al., 2017). Two-factor theorists sometimes posit that the first factor results from a specific neurological impairment (e.g., disconnection between autonomic signaling and face recognition in Capgras syndrome) rather than a general increase in noise, which would be expected to produce a much wider variety of abnormal experiences.

Adversarially learned inference provides a different perspective on these issues. Abnormal content arises from defects in the generator, which cause it to produce simulated data $\hat{x}$ and simulated interpretations $\hat{z}$ that have low probability under $P(x, z)$. These simulations are accepted by delusional patients because those patients also have a defect in their discriminator that impairs its ability to tell apart true and simulated samples. Thus, adversarially learned inference can be considered similar to two-factor theory, in the sense that it posits distinct impairments of abnormal content and abnormal evaluation.

The generative adversarial perspective offers a way to correct some of the shortcomings of prior Bayesian accounts. First, it suggests a broad hypothesis about the origin of delusional content (via an abnormal generator), whereas Bayesian models are silent on the origin of delusional content beyond the postulate that prediction errors are noisy. As discussed above,

noisy prediction errors seem inadequate to account for both the magnitude and specificity of delusional content. Second, the discriminator directly formalizes ideas about reality monitoring that have been applied to delusions, hallucinations, and confabulations (Bentall et al., 1991; Turner and Coltheart, 2010). In contrast, Bayesian models do not typically postulate any kind of specialized reality monitoring mechanism. While we have focused on delusions, the adversarial account may provide a broader framework that accompanies other kinds of reality distortion like hallucinations. The fact that hallucinations and delusions covary in schizophrenia (Grube et al., 1998) suggests that there may be a common underlying etiology.

6. DISCUSSION

This paper has assembled evidence across several disparate domains (perceptual phenomenology, neurobiology, and neuropsychiatry) in favor of a generative adversarial framework for approximate inference. In closing, we consider some broader issues and open questions.

6.1. Learning From the Imagination

Adversarially learned inference uses imagination to drive learning, exemplifying a broader class of imagination-based learning models that have been studied in cognitive science. The effects of imagination on learning have been widely documented (see Kappes and Morewedge, 2016, for a review). For example, Tartaglia et al. (2009) demonstrated that perceptual learning can occur through mental imagery, and related results have been observed across many different cognitive and behavioral tasks (Driskell et al., 1994; Gershman et al., 2017). It is unlikely that all imagination-based learning phenomena can be subsumed by the generative adversarial perspective. There are many ways that imagination could be involved in learning that don't involve adversarial interactions between a generator and a discriminator. For example, Niyogi et al. (1998) described how to use image transformations to produce "virtual examples" that can be used as additional training data, and Sutton (1990) developed related ideas for reinforcement learning. Both of these examples are forms of *data augmentation*, a technique widely used in machine learning to improve performance when data are limited (for some recent examples, see Hauberg et al., 2016; Ratner et al., 2017). Interestingly, generative adversarial algorithms have also ben employed for this purpose (Antoniou et al., 2017).

A key assumption of data augmentation algorithms is that the augmented data share certain properties with the true data distribution. In supervised learning, the augmented data must have the same labels as the true data. For example, Niyogi's technique is based on the idea that rigidly defined objects are invariant to rotations and translations. In reinforcement learning, augmented rewards and state transitions can be sampled from a learned model of the environment, as in Sutton's technique. The challenge, then, is to devise a scheme for producing augmented data with the right properties. Adversarially learned inference can be understood as one particular approach to this problem. The generator is not learning directly from the data distribution, but rather from a supervised signal (discriminator inaccuracy) that tells the generator how convincingly it has emulated the data distribution.

6.2. Toward a Synthesis of Approximate Inference Algorithms

Another broad issue concerns how we should make sense, and perhaps bring together, the menagerie of ideas about approximate inference in the brain. Adversarially learned inference shares elements of both Monte Carlo and variational algorithms. It uses samples to approximate expectations (as in Monte Carlo algorithms). But it also optimizes an objective function (the Jensen-Shannon divergence) that is closely related to standard variational algorithms (see Nowozin et al., 2016). Some generative adversarial approaches to inference make the connection even more explicit (Huszár, 2017; Mescheder et al., 2017). An interesting direction for future work will be to see whether some more systematic synthesis of these ideas is possible.

6.3. Predictions

Generative adversarial approaches to inference make a number of testable predictions. One is that impairment in the discriminator should lead to systematic distortions in learning, since imagined stimuli will be treated as real data. This should lead to generators that produce unrealistic samples, which could be tested by studying statistical learning in patients with prefrontal damage or with schizophrenia.

More broadly, the neural networks that have been developed for artificial intelligence tasks are designed to operate on high-dimensional data like natural images and videos, which opens up the possibility to make predictions about reality monitoring and subjective experience for real-world sensory inputs. For example, one could use them to predict which images are more likely to produce reality monitoring errors or meta-cognitive illusions in the periphery.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216, and by a research fellowship from the Alfred P. Sloan foundation.

ACKNOWLEDGMENTS

I am grateful to Talia Konkle, David Cox, Phil Corlett, Hakwan Lau, and George Alvarez for helpful discussions.

REFERENCES

Antoniou, A., Storkey, A., and Edwards, H. (2017). Data augmentation generative adversarial networks. *[Preprint]. arXiv:1711.04340.*

Beck, J., Pouget, A., and Heller, K. A. (2012). "Complex inference in neural circuits with probabilistic population codes and topic models," in *Advances in Neural Information Processing Systems*, 3059–3067.

Bentall, R., and Slade, P. D. (1985). Reality testing and auditory hallucinations: a signal detection analysis. *Br. J. Clin. Psychol.* 24, 159–169.

Bentall, R. P., Baker, G. A., and Havers, S. (1991). Reality monitoring and psychotic hallucinations. *Br. J. Clin. Psychol.* 30, 213–222.

Buda, M., Fornito, A., Bergström, Z. M., and Simons, J. S. (2011). A specific brain structural basis for individual differences in reality monitoring. *J. Neurosci.* 31, 14308–14313. doi: 10.1523/JNEUROSCI.3595-11.2011

Buesing, L., Bill, J., Nessler, B., and Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput. Biol.* 7:e1002211. doi: 10.1371/journal.pcbi.1002211

Chater, N. (2018). *The Mind is Flat: The Illusion of Mental Depth and the Improvised Mind*. London, UK: Penguin UK.

Collicutt, J., and Hemsley, D. (1981). A psychophysical investigation of auditory functioning in schizophrenia. *Br. J. Clin. Psychol.* 20, 199–204.

Coltheart, M., Menzies, P., and Sutton, J. (2010). Abductive inference and delusional belief. *Cogn. Neuropsyc.* 15, 261–287. doi: 10.1080/13546800903439120

Corlett, P. R., Frith, C. D., and Fletcher, P. C. (2009). From drugs to deprivation: a Bayesian framework for understanding models of psychosis. *Psychopharmacology* 206, 515–530. doi: 10.1007/s00213-009-1561-0

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.* 25, 410–418. doi: 10.1016/j.tree.2010.04.001

Dasgupta, I., Schulz, E., and Gershman, S. J. (2017). Where do hypotheses come from? *Cogn. Psychol.* 96, 1–25. doi: 10.1016/j.cogpsych.2017.05.001

Dasgupta, I., Schulz, E., Tenenbaum, J. B., and Gershman, S. J. (2019). A theory of learning to infer. *BioRxiv* 644534. doi: 10.1101/644534

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz machine. *Neural Comput.* 7, 889–904.

De Weerd, P. (2006). Perceptual filling-in: more than the eye can see. *Progress Brain Res.* 154, 227–245. doi: 10.1016/S0079-6123(06)54012-9

Dehaene, S., Charles, L., King, J.-R., and Marti, S. (2014). Toward a computational theory of conscious processing. *Curr. Opin. Neurobiol.* 25, 76–84. doi: 10.1016/j.conb.2013.12.005

Dennett, D. (1992). "Filling in versus' finding out: a ubiquitous confusion in cognitive science," in *Cognition, Conception, and Methodological Issues*, eds H. L. Pick Jr, P. van den Broek, and D. C. Knill (Washington, DC: American Psychological Association).

Diggle, P. J., and Gratton, R. J. (1984). Monte carlo methods of inference for implicit statistical models. *J. R. Stat. Soc. Ser. B* 46, 193–212.

Dinstein, I., Heeger, D. J., Lorenzi, L., Minshew, N. J., Malach, R., and Behrmann, M. (2012). Unreliable evoked responses in autism. *Neuron* 75, 981–991.

Donahue, J., Krähenbühl, P., and Darrell, T. (2016). "Adversarial feature learning," in *International Conference on Learning Representations*.

Driskell, J. E., Copper, C., and Moran, A. (1994). Does mental practice enhance performance? *J. Appl. Psychol.* 79, 481–492.

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., et al. (2017). "Adversarially learned inference," in *International Conference on Learning Representations*.

Farah, M. (1985). Psychophysical evidence for a shared representational medium for mental images and percepts. *J. Exp. Psychol. Gen.* 114, 91–103.

Farah, M. J., and Smith, A. F. (1983). Perceptual interference and facilitation with auditory imagery. *Percept. Psychophys.* 33, 475–478.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput. Biol.* 4:e1000211.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Garrison, J. R., Fernandez-Egea, E., Zaman, R., Agius, M., and Simons, J. S. (2017). Reality monitoring impairment in schizophrenia reflects specific prefrontal cortex dysfunction. *NeuroImage Clin.* 14, 260–268. doi: 10.1016/j.nicl.2017.01.028

Gershman, S. J., and Beck, J. M. (2017). "Complex probabilistic inference," in *Computational Models of Brain and Behavior*, ed A. Moustafa (Hoboken, NJ: Wiley-Blackwell).

Gershman, S. J., Zhou, J., and Kommers, C. (2017). Imaginative reinforcement learning: computational principles and neural mechanisms. *J. Cogn. Neurosci.* 29, 2103–2113. doi: 10.1162/jocn_a_01170

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2672–2680.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 290, 181–197.

Grube, B. S., Bilder, R. M., and Goldman, R. S. (1998). Meta-analysis of symptom factors in schizophrenia. *Schizoph. Res.* 31, 113–120.

Gutmann, M. U., and Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *J. Mach. Learn.Res.* 17, 4256–4302.

Haefner, R. M., Berkes, P., and Fiser, J. (2016). Perceptual decision-making as probabilistic inference by neural sampling. *Neuron* 90, 649–660. doi: 10.1016/j.neuron.2016.03.020

Hartig, F., Calabrese, J. M., Reineking, B., Wiegand, T., and Huth, A. (2011). Statistical inference for stochastic simulation models–theory and application. *Ecol. Lett.* 14, 816–827. doi: 10.1111/j.1461-0248.2011.01640.x

Hauberg, S., Freifeld, O., Larsen, A. B. L., Fisher, J., and Hansen, L. (2016). "Dreaming more data: class-dependent distributions over diffeomorphisms for learned data augmentation," in *Artificial Intelligence and Statistics*, 342–350.

Hemsley, D. R., and Garety, P. A. (1986). The formation of maintenance of delusions: a Bayesian analysis. *Br. J. Psychiatry* 149, 51–56.

Huang, Y., and Rao, R. P. (2011). Predictive coding. *Wiley Interdiscipl. Rev. Cogn. Sci.* 2, 580–593. doi: 10.1002/wcs.142

Huszár, F. (2017). Variational inference using implicit distributions. *arXiv preprint* arXiv:1702.08235.

Ishai, A., and Sagi, D. (1995). Common mechanisms of visual imagery and perception. *Science* 268, 1772–1774.

Johnson, M., and Raye, C. (1981). Reality monitoring. *Psychol. Rev.* 88, 67–85.

Kappes, H. B., and Morewedge, C. K. (2016). Mental simulation as substitute for experience. *Soc. Personal. Psychol. Compass* 10, 405–420. doi: 10.1111/spc3.12257

Kensinger, E., and Schacter, D. (2006). Neural processes underlying memory attribution on a reality-monitoring task. *Cereb. Cortex* 16, 1126–1133. doi: 10.1093/cercor/bhj054

Knill, D. C., and Richards, W. (1996). *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.

Lau, H. (2019). Consciousness, metacognition, & perceptual reality monitoring. *PsyArXiv.* doi: 10.31234/osf.io/ckbyf

Lau, H., and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* 15, 365–373. doi: 10.1016/j.tics.2011.05.009

Lee, T. S., and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am.* 20, 1434–1448. doi: 10.1364/JOSAA.20.001434

Lochmann, T., and Deneve, S. (2011). Neural processing as causal inference. *Curr. Opin. Neurobiol.* 21, 774–781. doi: 10.1016/j.conb.2011.05.018

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat. Neurosci.* 9:1432. doi: 10.1038/nn1790

McKay, R. (2012). Delusional inference. *Mind Lang.* 27, 330–355. doi: 10.1111/j.1468-0017.2012.01447.x

Mescheder, L., Nowozin, S., and Geiger, A. (2017). "Adversarial variational bayes: unifying variational autoencoders and generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (JMLR. org.), 2391–2400.

Niyogi, P., Girosi, F., and Poggio, T. (1998). Incorporating prior information in machine learning by creating virtual examples. *Proceed. IEEE* 86, 2196–2209.

Noë, A., Pessoa, L., and Thompson, E. (2000). Beyond the grand illusion: what change blindness really teaches us about vision. *Visual Cogn.* 7, 93–106. doi: 10.1080/135062800394702

Nowozin, S., Cseke, B., and Tomioka, R. (2016). "f-gan: training generative neural samplers using variational divergence minimization," in *Advances in Neural Information Processing Systems*, 271–279.

Odegaard, B., Chang, M. Y., Lau, H., and Cheung, S.-H. (2018). Inflation versus filling-in: why we feel we see more than we actually do in peripheral vision. *Philos. Trans. R. Soc. B Biol. Sci.* 373:20170345. doi: 10.1098/rstb.2017.0345

Orbán, G., Berkes, P., Fiser, J., and Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* 92, 530–543. doi: 10.1016/j.neuron.2016.09.038

Park, W. J., Schauder, K. B., Zhang, R., Bennetto, L., and Tadin, D. (2017). High internal noise and poor external noise filtering characterize perception in autism spectrum disorder. *Sci. Rep.* 7:17584. doi: 10.1038/s41598-017-17676-5

Perky, C. W. (1910). An experimental study of imagination. *Am. J. Psychol.* 21, 422–452.

Raju, R. V., and Pitkow, Z. (2016). "Inference by reparameterization in neural population codes," in *Advances in Neural Information Processing Systems*, 2029–2037.

Ramachandran, V. S., Blakeslee, S., and Shah, N. (1998). *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. New York, NY: William Morrow.

Ramachandran, V. S., and Hirstein, W. (1997). Three laws of qualia: what neurology tells us about the biological functions of consciousness. *J. Consci. Stud.* 4, 429–457.

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87.

Ratner, A. J., Ehrenberg, H., Hussain, Z., Dunnmon, J., and Ré, C. (2017). "Learning to compose domain-specific transformations for data augmentation," in *Advances in Neural Information Processing Systems*, 3236–3246.

Sanborn, A. N., and Chater, N. (2016). Bayesian brains without probabilities. *Trends Cogn. Sci.* 20, 883–893. doi: 10.1016/j.tics.2016.10.003

Sanborn, A. N., and Silva, R. (2013). Constraining bridges between levels of analysis: a computational justification for locally bayesian learning. *J. Math. Psychol.* 57, 94–106. doi: 10.1016/j.jmp.2013.05.002

Segal, S., and Fusella, V. (1970). Influence of imaged pictures and sounds on detection of visual and auditory signals. *J. Exp. Psychol. Gen.* 83, 458–464. doi: 10.1037/h0028840

Seth, A. K., Suzuki, K., and Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Front. Psychol.* 2:395. doi: 10.3389/fpsyg.2011.00395

Simons, D. J. (2000). Current approaches to change blindness. *Visual Cogn.* 7, 1–15. doi: 10.1080/135062800394658

Simons, J. S., Garrison, J. R., and Johnson, M. K. (2017). Brain mechanisms of reality monitoring. *Trends Cogn. Sci.* 21, 462–473. doi: 10.1016/j.tics.2017.03.012

Simons, J. S., Henson, R. N., Gilbert, S. J., and Fletcher, P. C. (2008). Separable forms of reality monitoring supported by anterior prefrontal cortex. *J. Cogn. Neurosci.* 20, 447–457. doi: 10.1162/jocn.2008.20.3.447

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., et al. (2018). The predictive coding account of psychosis. *Biol. Psychiatry* 84, 634–643. doi: 10.1016/j.biopsych.2018.05.015

Sutton, R. S. (1990). "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in *Machine Learning Proceedings* (Elsevier), 216–224.

Tartaglia, E. M., Bamert, L., Mast, F. W., and Herzog, M. H. (2009). Human perceptual learning by mental imagery. *Curr. Biol.* 19, 2081–2085. doi: 10.1016/j.cub.2009.10.060

Tranel, D., Damasio, H., and Damasio, A. R. (1995). Double dissociation between overt and covert face recognition. *J. Cogn. Neurosci.* 7, 425–432.

Turner, M., and Coltheart, M. (2010). Confabulation and delusion: a common monitoring framework. *Cogn. Neuropsych.* 15, 346–376. doi: 10.1080/13546800903441902

van Schalkwyk, G. I., Volkmar, F. R., and Corlett, P. R. (2017). A predictive coding account of psychotic symptoms in autism spectrum disorder. *J. Autism Dev. Disord.* 47, 1323–1340. doi: 10.1007/s10803-017-3065-9

Vuilleumier, P., Mohr, C., Valenza, N., Wetzel, C., and Landis, T. (2003). Hyperfamiliarity for unknown faces after left lateral temporo-occipital venous infarction: a double dissociation with prosopagnosia. *Brain* 126, 889–907. doi: 10.1093/brain/awg086

Winterer, G., and Weinberger, D. R. (2004). Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends Neurosci.* 27, 683–690. doi: 10.1016/j.tins.2004.08.002

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 21 January 2020
doi: 10.3389/frai.2019.00031

Analysis of Features Selected by a Deep Learning Model for Differential Treatment Selection in Depression

Joseph Mehltretter[1†], *Colleen Rollins*[2†], *David Benrimoh*[3,4,5,6*], *Robert Fratila*[6], *Kelly Perlman*[5,6], *Sonia Israel*[5,6], *Marc Miresco*[6,7], *Marina Wakid*[5] *and Gustavo Turecki*[3,5]

[1] *Department of Computer Science, University of Southern California, Los Angeles, CA, United States,* [2] *Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom,* [3] *Department of Psychiatry, McGill University, Montreal, QC, Canada,* [4] *Faculty of Medicine, McGill University, Montreal, QC, Canada,* [5] *Douglas Mental Health University Institute, Montreal, QC, Canada,* [6] *Aifred Health, Montreal, QC, Canada,* [7] *Department of Psychiatry, Jewish General Hospital, Montreal, QC, Canada*

Background: Deep learning has utility in predicting differential antidepressant treatment response among patients with major depressive disorder, yet there remains a paucity of research describing how to interpret deep learning models in a clinically or etiologically meaningful way. In this paper, we describe methods for analyzing deep learning models of clinical and demographic psychiatric data, using our recent work on a deep learning model of STAR*D and CO-MED remission prediction.

Methods: Our deep learning analysis with STAR*D and CO-MED yielded four models that predicted response to the four treatments used across the two datasets. Here, we use classical statistics and simple data representations to improve interpretability of the features output by our deep learning model and provide finer grained understanding of their clinical and etiological significance. Specifically, we use representations derived from our model to yield features predicting both treatment non-response and differential treatment response to four standard antidepressants, and use linear regression and *t*-tests to address questions about the contribution of trauma, education, and somatic symptoms to our models.

Results: Traditional statistics were able to probe the input features of our deep learning models, reproducing results from previous research, while providing novel insights into depression causes and treatments. We found that specific features were predictive of treatment response, and were able to break these down by treatment and non-response categories; that specific trauma indices were differentially predictive of baseline depression severity; that somatic symptoms were significantly different between males and females, and that education and low income proved important psycho-social stressors associated with depression.

Conclusion: Traditional statistics can augment interpretation of deep learning models. Such interpretation can lend us new hypotheses about depression and contribute to building causal models of etiology and prognosis. We discuss dataset-specific effects and ideal clinical samples for machine learning analysis aimed at improving tools to assist in optimizing treatment.

Keywords: deep learning, features, depression, interpretability, treatment

OPEN ACCESS

Edited by:
Casper Hesp, Amsterdam Brain and Cognition (ABC), Netherlands

Reviewed by:
Maria Chan, Memorial Sloan Kettering Cancer Center, United States
Robert Rallo, Pacific Northwest National Laboratory, United States

***Correspondence:**
David Benrimoh
david.benrimoh@mail.mcgill.ca

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *30 October 2019*
Accepted: *06 December 2019*
Published: *21 January 2020*

Citation:
Mehltretter J, Rollins C, Benrimoh D, Fratila R, Perlman K, Israel S, Miresco M, Wakid M and Turecki G (2020) Analysis of Features Selected by a Deep Learning Model for Differential Treatment Selection in Depression. Front. Artif. Intell. 2:31. doi: 10.3389/frai.2019.00031

INTRODUCTION

The heterogeneity of depression constitutes a major barrier to successful treatment (Perna et al., 2018). Clinicians and patients are faced with a plethora of treatment options, with over 20 commonly prescribed antidepressants, augmentation therapies, psychotherapies, neuromodulation, and lifestyle interventions, but a paucity of evidence-based information to inform treatment selection and personalization. The resultant trial and error approach to treatment selection prescription is ineffective: a third of patients fail to remit to a first-line antidepressant, with remission rates decreasing with subsequent treatments (Rush et al., 2006). Researchers have strived to identify predictors of treatment outcome across clinical profile, sociodemographic, physiological, neuroimaging, genomic, and other possible predictor types (Williams et al., 2011), yet few, if any, predictors have translated into common clinical practice. Machine learning (ML) is capable of tackling the challenges of interpreting large, multidimensional, interrelated datasets found in psychiatric research and may help us create clinically useful models for treatment selection.

Two objectives in the study of biological systems are inference and prediction. Inference creates a model of data-generation to test a hypothesis about how a particular system behaves, whereas prediction forecasts possible outcome or behavior without necessarily understanding underlying biological mechanisms (Bzdok et al., 2018). Classical statistical methods, such as regression and *t*-tests, focus on inference and have been a dominant method for analyzing psychiatric data and offering insight into causal associations. For instance, logistic regression models assessing the association of demographic and clinical characteristics on treatment outcome in the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial, a large multicenter sequenced treatment trial for depression, have shown that race, low education, post-traumatic stress disorder (PTSD), and hypochondriasis are independently associated with worsened depression (Friedman et al., 2009), as well as depression severity, energy/fatigue, race, education, and PTSD occurrence (Perlis, 2013); in addition, having witnessed or experienced trauma has been used to estimate risk for treatment-resistance among major depressive disorder (MDD) outpatients (Perlis, 2013). These results are bolstered with receiver operating characteristic (ROC) analyses also showing income and education to be predictors of response in STAR*D (Jakubovski and Bloch, 2014). However, in recent years, classical statistics and null hypothesis significance testing frameworks have been increasingly scrutinized due to the emphasis on *p*-value testing and difficulties with reproducibility (Wagenmakers, 2007). In contrast, machine learning allows for individualized prediction through the implementation of learning algorithms, which make fewer assumptions about data-generation, to find patterns in large, heterogeneous datasets. Advances in machine learning have highlighted its utility in identifying patterns in complex data for psychiatric research (Iniesta et al., 2016; Passos et al., 2016) and specifically for outcomes of depression treatments (Lee et al., 2018). Recent studies have leveraged machine learning methods to predict antidepressant treatment response for individuals with depression, identifying 25 features most predictive of whether a patient will respond to citalopram (Chekroud et al., 2016), predicting persistence, chronicity, and severity of depression from self-report questionnaires (Kessler et al., 2016), predicting treatment response to electroconvulsive therapy (ECT) using baseline hippocampal subfield volumes (Cao et al., 2018), predicting treatment resistance before initiation of a second antidepressant (Nie et al., 2018), using deep learning to predict response to SSRIs (Lin et al., 2018), and using Random Forests to predict outcome in treatment-resistant depression (Kautzky et al., 2018). However, the non-linearity of relationships that ML techniques capture in models make it difficult to integrate ML with existing biological knowledge and clinical practice, where researchers, clinicians, and patients often seek to understand causal relationships. We suggest that deep learning and traditional statistics can be used in a complementary fashion to interpret clinically meaningful associations.

One goal of personalized psychiatry is to predict a given patient's pre-treatment likelihood of response to an array of treatments in order to aid in selecting the treatment with the highest likelihood of response before therapy is administered. In recent work (Mehltretter et al., 2019), we performed a deep learning analysis on the Combining medications to enhance depression outcomes (CO-MED) clinical trial and Level 1 of STAR*D. Of all machine learning techniques, deep learning is considered one of the most effective, but also the most difficult to interpret (Zhang et al., 2018). We produced an algorithm that predicts response to four antidepressant treatments and is theoretically capable of increasing population remission rates via differential treatment benefit prediction (Mehltretter et al., 2019). Our study yielded four models, described below. As we examined each model's features found to be most predictive of remission, we identified striking consistencies in the features across models, and between our work and that of Chekroud et al. (2016) and others, as well as some surprising inconsistencies. Improving interpretability of deep learning models is important for translational research and for increasing their clinical utility. In our previous paper, we produced "interpretability reports" that helped understand the key features for predictions for individual patients. In this paper, we use regression and classical statistics to help interpret our results in order to better understand what complex ML outputs can tell us about the mechanisms driving remission to depression, and the relationships between predictive features. Based on these observations, we ask clinically- and mechanistically-relevant questions concerning general vs. specific predictors of response to antidepressants, trauma-related features, dataset differences in education, somatic symptoms and gender, using simple data representations and manipulations and traditional statistics, such as regression and *t*-tests. We evaluate our findings in the context of existing hypotheses concerning the etiology and prognosis of major depression and use what we learn to offer new directions for depression research and the use of ML in psychiatric data science.

TABLE 1 | Optimal features selected by the deep learning algorithm for remission prediction.

		Model (number of features)				
Category		**Combined (17)**	**STAR*D optimal (21)**	**STAR*D tested on CO-MED (14)**	**CO-MED alone (26)**	**Chekroud et al. (2016) (25)**
Sociodemographic		Number of years in formal education	Number of years in formal education	Number of years in formal education		Years of education
		Monthly household income	Monthly household income	Monthly household income		
						Black or African American
						White
			Current marital status			
			Months lived at residence			
			Has private insurance			
Patient history			Patient has a history of psychotropic meds		Previously taken zoloft sertraline	Ever taken sertraline
					Previously taken Prozac fluoxetine	
			Child history of depression			Number of previous major depressive episodes
Symptom profile (depression)	Depression severity	Initial QIDS total severity	Initial QIDS total severity	Initial QIDS total severity		Initial QIDS total severity
			Initial HAM-D depression severity			Initial HAM-D depression severity
		HAM-D suicide	QIDS suicidal ideation	QIDS suicidal ideation	QIDS suicidal ideation	HAM-D suicide
					Past 2 weeks: Considered hurting self or wished they were dead	
		QIDS mood (sad)	QIDS mood (sad)	QIDS mood (sad)		QIDS mood (sad)
						Depressed mood most of the day, nearly every day
	Somatic	HAM-D somatic energy		HAM-D somatic energy		HAM-D somatic energy
						HAM-D somatic anxiety
		Have you ever been bothered by aches and pains in many different parts of your body?	Have you ever been bothered by aches and pains in many different parts of your body?	Have you ever been bothered by aches and pains in many different parts of your body?	Have you ever been bothered by aches and pains in many different parts of your body?	Have you ever been bothered by aches and pains in many different parts of your body?
		QIDS weight (increase) last 2 weeks		QIDS weight (increase) last 2 weeks	Dysthymic disorder/major depressive episode. Weight loss or weight gain or appetite change	
		Eat a lot when not hungry			Feel disgusted after overeating	
	Sleep	QIDS sleep onset insomnia			Sleep onset insomnia	QIDS sleep onset insomnia
					I have been having more trouble sleeping than usual	HAM-D delayed insomnia

(Continued)

TABLE 1 | Continued

Category		Model (number of features)				
		Combined (17)	**STAR*D optimal (21)**	**STAR*D tested on CO-MED (14)**	**CO-MED alone (26)**	**Chekroud et al. (2016) (25)**
		QIDS energy or fatigability				QIDS energy or fatigability
	Cognitive or behavioral					QIDS psychomotor agitation
					QIDS concentration/decision making	
					Dysthymic disorder/major depressive episode: Poor concentration or difficulty making decisions	
						HAM-D loss of insight
					Feelings of worthlessness or guilt	
Comorbidity: Trauma		Have you ever witnessed a traumatic event such as rape, assault, someone dying in an accident, or any other extremely upsetting event?	Have you ever witnessed a traumatic event such as rape, assault, someone dying in an accident, or any other extremely upsetting event?	Have you ever witnessed a traumatic event such as rape, assault, someone dying in an accident, or any other extremely upsetting event?		Have you ever witnessed a traumatic event such as rape, assault, someone dying in an accident, or any other extremely upsetting event?
					Avoid activities that remind you of trauma	Did you try to avoid activities, places, or people that reminded you of a traumatic event?
	Jumpy because of a trauma	Jumpy because of a trauma	Jumpy because of a trauma			
		Did reminders of a traumatic event make you shake, break out into a sweat, or have a racing heart?				Did reminders of a traumatic event make you shake, break out into a sweat, or have a racing heart?
			Axis I: Post-traumatic stress disorder			
				Feel distant because of trauma		
Comorbidity: Anxiety		Anxiety being in crowded places	Anxiety being in crowded places	Anxiety being in crowded places		
			Did any of the following make you feel fearful, anxious, or nervous because you were afraid you'd have an anxiety attack in the situation? Standing in long lines	Did any of the following make you feel fearful, anxious, or nervous because you were afraid you'd have an anxiety attack in the situation? Standing in long lines		Did any of the following make you feel fearful, anxious, or nervous because you were afraid you'd have an anxiety attack in the situation? Standing in long lines

(Continued)

TABLE 1 | Continued

Category		Model (number of features)				
		Combined (17)	**STAR*D optimal (21)**	**STAR*D tested on CO-MED (14)**	**CO-MED alone (26)**	**Chekroud et al. (2016) (25)**
						Did any of the following make you feel fearful, anxious, or nervous because you were afraid you'd have an anxiety attack in the situation? Driving or riding in a car
					Avoid situation because afraid of anxiety attack	Did you have attacks of anxiety that caused you to avoid certain situations or to change your behavior or normal routine?
					Anxiety attacks for no reason	
Function			Current employment status			Currently employed
	How many hours did you actually work	How many hours did you actually work	How many hours did you actually work			
Symptom profile: Other			Neurological			
			Lower gastrointestinal (GI)			
					I talk more than usual	
					I suddenly feel very confident	
					I can feel my heart racing	
					Worry about saying something stupid	
					Worry about embarrassing self	
					Worry something you forgot	
					Guilt feelings and delusions	
					Hallucinations	
					Sleep disturbance	
Miscellaneous		Drug assigned			Assigned to randomization arm	

*This table demonstrates the features composing each studied model. Note: for trauma, the following features were found to be predictive in STAR*D and CO-MED: "jumpy because of a trauma," "ever witnessed a traumatic event," and "Did reminders of a traumatic event make you shake, break out into a sweat, or have a racing heart?".*

MATERIALS AND METHODS

Here we discuss the data and models produced as part of our previous analysis (Mehltretter et al., 2019). We provide detailed methods in the **Supplementary Methods** section.

Datasets

Data from CO-MED [Combining Medications to Enhance Depression Outcomes (COMED); ClinicalTrials.gov, NCT00590863] and STAR*D Level 1 (STAR*D; ClinicalTrials.gov, NCT00021528) were used for these analyses.

TABLE 2 | Ten-fold cross validated model accuracy metrics.

Model (Number of features)	AUC	NPV	PPV	Sensitivity	Specificity
Combined STAR*D + CO-MED (17)	0.69	0.64	0.64	0.60	0.60
STAR*D Optimal (21)	0.71	0.68	0.68	0.69	0.69
STAR*D Model that was then tested on CO-MED (14)	0.70	0.64	0.64	0.60	0.60
CO-MED Alone (26)	0.80	0.64	0.64	0.60	0.60
Chekroud et al. (2016) (STAR*D only) (25)	0.70	0.65	0.64	0.63	0.66

*AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value; STAR*D, sequenced treatment alternatives to relieve depression; CO-MED, combining medications to enhance depression outcomes.*

CO-MED enrolled 665 outpatients who were randomly assigned three possible treatments: escitalopram and placebo, bupropion and escitalopram, or mirtazapine and venlafaxine. STAR*D Level 1 enrolled 2,757 subjects, all of whom were treated with citalopram.

Feature Selection

A feature selection and analysis pipeline was used that consisted of variance thresholding, recursive feature elimination with cross validation, and feature importance extraction using a randomized lasso algorithm. The parameters for each method were optimized by analyzing the accuracy of the neural network's predictions about remission. Full details can be found in Mehltretter et al. (2019).

Neural Network

A dense neural network was built with Vulcan (https://github.com/Aifred-Health/Vulcan) to train and evaluate our remission prediction capabilities. Since our data were limited in dimensionality we configured our neural networks to prevent over fitting by using a more shallow network. Each node within the network used scaled exponential linear unit function for activation, and softmax was used on the final layer for predicting the probability of remission.

Models

We produced four models from different combinations of features from STAR*D and CO-MED, and compared these to a previously published model, and they are as follows:

(1) **Combined model:** The combined model was developed by merging the STAR*D dataset (2,757 subjects, 1 treatment group) with the CO-MED dataset (665 subjects, 3 treatment groups) and removing features that were not common to both datasets, resulting in 3,222 patients, 4 treatment groups, and 213 features. We used variance thresholding and recursive feature elimination with cross validation to determine the features most salient for differential treatment prediction. This procedure identified 17 features.

(2) **STAR*D Optimal model:** This remission-prediction model was trained on the citalopram data from level 1 of STAR*D, including all possible features in STAR*D without eliminating those not found in CO-MED, and was then validated using internal cross-validation.

(3) **STAR*D Tested on CO-MED:** This model predicted remission with citalopram using features common to STAR*D and CO-MED, and generalized to the three branches of CO-MED to ensure our model wasn't biased toward citalopram.

(4) **CO-MED Alone:** This model predicts remission for within the CO-MED dataset alone across all drug categories, including all the features present in CO-MED before feature selection. Six hundred and sixty five subjects were included and 25 features were used after feature selection for predicting remission.

(5) **Chekroud et al. (2016) model:** We include results from the model detailed in Chekroud et al. (2016) to allow for direct comparison to our models. Chekroud et al. (2016) trained a gradient-boosting model on the citalopram data from level 1 of the STAR*D dataset and tested it on the three treatment groups of the CO-MED dataset, producing 25 features.

Table 1 demonstrates the features selected by the deep learning algorithm. Model performance metrics are reported in **Table 2**.

Interpretation of Model Features

We set out to understand the features in these models and how they might relate to mechanisms of response in depression treatment and determination of initial depression severity, as this is an important predictor of response to treatment. We outline key observations from **Table 1** that motivated five specific questions:

(1) **Predictors of remission vs. predictors of response to specific antidepressants**

By combining data from the STAR*D and CO-MED clinical trials for a pooled dataset across 4 treatments, we present a model that is able to perform differential treatment prediction. A benefit of this contribution is that we can begin to disentangle features that are predictive of remission regardless of drug category from features that are predictive of remission to specific drugs. We observed that two features were predictive of remission across all 5 models (**Table 1**): "Have you ever been bothered by aches and pains in many different parts of your body?" and suicidal ideation score. Their commonality across all models suggests that these are *general* predictors of response to

antidepressant treatment, which reproduces some results from extant literature, in which suicidal ideation and somatic symptoms are robust contributors to more severe course of illness, increased rates of relapse, higher risk of suicide, and greater burden of care (Papakostas et al., 2003; Kapfhammer, 2006; Bohman et al., 2012). Four features—Number of years of formal education (beginning at grade 1), having witnessed a traumatic event initial depression severity [as assessed by Quick Inventory for Depressive Symptomatology (QIDS)], and sad mood (QIDS)—were common to all models except for the COMED-alone model. This suggests two non-mutually exclusive possibilities: that these represent citalopram-specific predictors of response, or that there were differences between the STAR*D and COMED samples, despite their large size and fairly broad inclusion criteria aimed at generating representative MDD samples. Given the possibility of antidepressant-specific vs. general predictors of response, we asked:

"Can we identify features predictive of response to each of the four antidepressants within our model (escitalopram, bupropion, venlafaxine-mirtazapine, citalopram) individually, as well as to the subgroup of patients with a low probability of responding to any of the drugs?"

(2) **Trauma-related features**

Specific indices of trauma emerged from the deep learning model as predictive of treatment response for both the STAR*D and COMED datasets. Since trauma is also a strong risk factor for depression onset and severity (Nelson et al., 2017), this led us to question:

"Are specific aspects of trauma predictive of baseline depression?"

(3) **Differences in education level between datasets**

While level of education was a feature that was relevant for predicting remission in STAR*D alone and in the combined dataset, it was not predictive in the CO-MED dataset alone. Since the combined dataset is biased toward STAR*D due to its larger sample size, this could explain the presence of the education feature the STAR*D and CO-MED combined dataset. We therefore analyzed the difference between levels of education for the two separate datasets to answer the question:

"Do the participants in STAR*D and CO-MED come from the same population, or are these populations different in key variables that are predictive of outcomes?"

(4) **Somatic symptoms and gender**

Each of the four deep learning models retained somatic symptoms of depression, such as feeling aches and pains, as being important predictors of remission (**Table 1**). Gender, however, was not selected as an optimal feature predictive of remission. This could indicate that our model was not concerned with gender because it was able to extract specific features that differed between genders and therefore did not need to use gender as a proxy. Given that somatic symptoms have previously been shown to differ by gender (Silverstein et al., 2013), we asked:

"Do somatic symptoms of depression differ by gender?"

Statistics

The data were analyzed at a Bonferroni-corrected significance level of $p < 0.005$ with the statistical software RStudio version 1.0.136. Statistical tests used were student's *t*-tests and linear regression.

RESULTS

Can We Identify Features Predictive of Response to Each of the Four Antidepressants Within Our Model (Escitalopram, Bupropion, Venlafaxine-Mirtazapin, Citalopram) Individually, and Features Suggestive of a Low Probability of Responding to Any of the Drugs?

Given the four possible medications within our model, we assessed which features were important for predicting remission [as defined by a score of 5 or less on the Quick Inventory of Depressive Symptomatology (QIDS)] for each individual drug, as well as which features were predictive of a low probability of remission with any drug. We first defined a low probability of remission to any of the drugs as being a patient whose remission probability for each drug was less than the baseline population remission rate. This resulted in five sub-groups: one group for each of the four treatments, and a fifth group with a low probability of remission to any treatment. We created a set of 750 subjects: 500 randomly selected from the STAR*D study and 250 subjects randomly selected from the CO-MED trial. We assigned subjects to a sub-group by running our test set of subjects through our trained model four times, each time with a new medication storing the probability of remission for that given subject with that medication. We were, in effect, generating potential outcomes under each of four different treatments to see whether a patient would be predicted to experience remission under any or none of drugs. We then assigned each subject in a group based on the medication that produced the highest probability of remission. If no drug had a remission probability of greater than the baseline remission rate (34%), the patient was assigned to the non-remission group. This produced the following group sizes (**Table 3**).

We then used saliency maps to identify the importance of each individual feature with regards to producing the given probability of remission, and took the top five for each subject. **Tables 4–8** show how often a feature was found to be in the top

TABLE 3 | Number of subjects in each subgroup.

Group	Number of subjects
Non-remission	373
Escitalopram	28
Escitalopram-bupropion	28
Venlafaxine-mirtazapine	53
Citalopram	268

TABLE 4 | Non-remission subgroup feature information.

Feature	% occurrence in top five
Initial QIDS total severity	13.19
Have you ever been bothered by aches and pains in many different parts of your body?	13
Number of years in formal education	11.84
HAM-D somatic energy	9.71
QIDS energy or fatigability	9.33
Eat a lot when not hungry	7.27
QIDS sleep onset insomnia	6.1
Monthly household income	5.68
QIDS mood (sad)	5.68
Have you ever witnessed a traumatic event such as rape, assault, someone dying in an accident, or any other extremely upsetting event?	3.97
Jumpy because of a trauma	2.63
Anxiety being in crowded places	1.82
How many hours did you actually work	0.97
QIDS weight (increase) last 2 weeks	0.86
HAM-D suicide	0.75

HAM-D, Hamilton Depression Rating Scale; QIDS, Quick Inventory of Depressive Symptomatology.

TABLE 5 | Escitalopram subgroup feature information.

Feature	% occurrence in top five
QIDS sleep onset insomnia	14.28
HAM-D somatic energy	13.57
Monthly household income	13.57
QIDS mood (sad)	13.57
Number of years in formal education	5.7
Jumpy because of a trauma	5
HAM-D suicide	5
How many hours did you actually work	5
Eat a lot when not hungry	1.43
QIDS energy or fatigability	1.43
Have you ever been bothered by aches and pains in many different parts of your body?	0.71
QIDS weight (increase) last 2 weeks	0.71
Initial QIDS total severity	0.71

HAM-D, Hamilton Depression Rating Scale; QIDS, Quick Inventory of Depressive Symptomatology.

TABLE 6 | Escitalopram bupropion subgroup feature information.

Feature	% occurrence in top five
HAM-D somatic energy	14.29
Monthly household income	14.29
QIDS sleep onset insomnia	14.29
QIDS mood (sad)	14.29
Number of years in formal education	5.7
Jumpy because of a trauma	4.29
HAM-D suicide	4.29
How many hours did you actually work	4.29
Eat a lot when not hungry	1.43
QIDS energy or fatigability	1.43
Have you ever been bothered by aches and pains in many different parts of your body?	1.43
Initial QIDS total severity	1.43

HAM-D, Hamilton Depression Rating Scale; QIDS, Quick Inventory of Depressive Symptomatology.

TABLE 7 | Venlafaxine-mirtazapine subgroup feature information.

Feature	% occurrence in top five
HAM-D somatic energy	14.33
Monthly household income	10.94
QIDS mood (sad)	10.94
QIDS sleep onset insomnia	10.57
Number of years in formal education	8.68
Initial QIDS total severity	6.8
Have you ever been bothered by aches and pains in many different parts of your body?	6.41
Have you ever witnessed a traumatic event such as rape, assault, someone dying in an accident, or any other extremely upsetting event?	3.77
Jumpy because of a trauma	3.01
How many hours did you actually work	3.01
QIDS energy or fatigability	2.64
HAM-D suicide	2.64
Eat a lot when not hungry	2.26
Anxiety being in crowded places	0.75

HAM-D, Hamilton Depression Rating Scale; QIDS, Quick Inventory of Depressive Symptomatology.

TABLE 8 | Citalopram subgroup feature information.

Feature	% occurrence in top five
HAM-D somatic energy	14.25
QIDS mood (sad)	10.15
Monthly household income	10.15
Number of years in formal education	9.6
Initial QIDS total severity	8.28
Have you ever been bothered by aches and pains in many different parts of your body?	8.21
Have you ever witnessed a traumatic event such as rape, assault, someone dying in an accident, or any other extremely upsetting event?	4.2
QIDS energy or fatigability	4.2
Eat a lot when not hungry	4.1
Jumpy because of a trauma	1.72
How many hours did you actually work	1.57
HAM-D suicide	1.5
Anxiety being in crowded places	0.22
QIDS weight (increase) last 2 weeks	0.15
Did reminders of a traumatic event make you shake, break out into a sweat, or have a racing heart?	0.07

HAM-D, Hamilton Depression Rating Scale; QIDS, Quick Inventory of Depressive Symptomatology.

five features for each sub-group, indicating the frequency, at the individual patient level, that this feature figured as one of the most influential features in the probability calculation.

Are Specific Aspects of Trauma Predictive of Baseline Depression?

We performed multiple linear regression analyses inputting the three trauma features deemed important by our deep learning

model as predictors to explore the relationship between trauma and baseline QIDS score. One model included "jumpy because of traumatic event," "witnessed traumatic event," "shaky because of trauma,"; a second model also included gender and years of education as covariates. The linear regression models showed that only "Did reminders of a traumatic event make you shake, break out into a sweat, or have a racing heart?" was significantly associated with baseline depression severity, an association that remained after controlling for gender and years of education. Gender and years of education were also significantly predictive of baseline QIDS score (**Table 9**).

TABLE 9 | Results of linear regression analyses examining the contribution of trauma indices to baseline depression severity in STAR*D.

	Model 1	Model 2
	Beta estimate (S.E.) [95% CI]	**Beta estimate (S.E.) [95% CI]**
Have you ever witnessed a traumatic event such as rape, assault, someone dying in an accident, or any other extremely upsetting event	0.246 (0.148) [−0.04, 0.54]	0.309 (0.171) [−0.03, 0.65]
Jumpy because of trauma	0.415 (0.177) [0.07, 0.76]	0.288 (0.20) [−0.108, 0.69]
Did reminders of a traumatic event make you shake, break out into a sweat, or have a racing heart?	1.027 (0.177) [0.68, 1.37]*	0.813 (0.202) [0.42–1.21]*
Gender		−0.794 (0.178) [−1.144, −0.44]*
Years of education		−0.098 (0.024) [−0.15, −0.05]*
F-statistic	30.00	17.97
N	2,696	1,951
R2	0.032	0.042

*$p < 0.005$, the cut-off determined via a Bonferroni correction.

Do the Participants in STAR*D and CO-MED Come From the Same Population, or Are These Populations Different in Key Variables That Are Predictive of Outcomes?

As education is more predictive of outcome in the STAR*D as compared to the CO-MED data, we performed independent *t*-tests to identify whether the distribution of education itself varied between participant samples, since education is unlikely to be a drug-specific predictor. As observed in **Figure 1**, a *t*-test showed there was no appreciable difference between the years of education in the CO-MED (orange bars) and STAR*D (blue bars) participants (mean difference = 0.06, $p = 0.678$).

Do Somatic Symptoms of Depression Differ by Gender?

We used *t*-tests to see if somatic symptoms of depression differed between the genders. **Table 10** details the difference in somatic symptoms between males and females, finding significant differences for the following features: somatic energy as measured by the Hamilton Depression Rating Scale (HAM-D), being bothered by aches/pains, and energy/fatigability, as measured by the Quick Inventory of Depressive Symptomatology (QIDS).

TABLE 10 | Significant differences in somatic symptoms of depression between males and females.

	Somatic energy	Bothered by aches/pains	Weight (increase) last 2 weeks	Energy/ fatigability
Male to Female Difference	−0.14*	−0.07*	−0.13	−0.21*
p-value	0.000	0.0001	0.027	0.000

*$p < 0.005$, the cut-off determined via a Bonferroni correction.

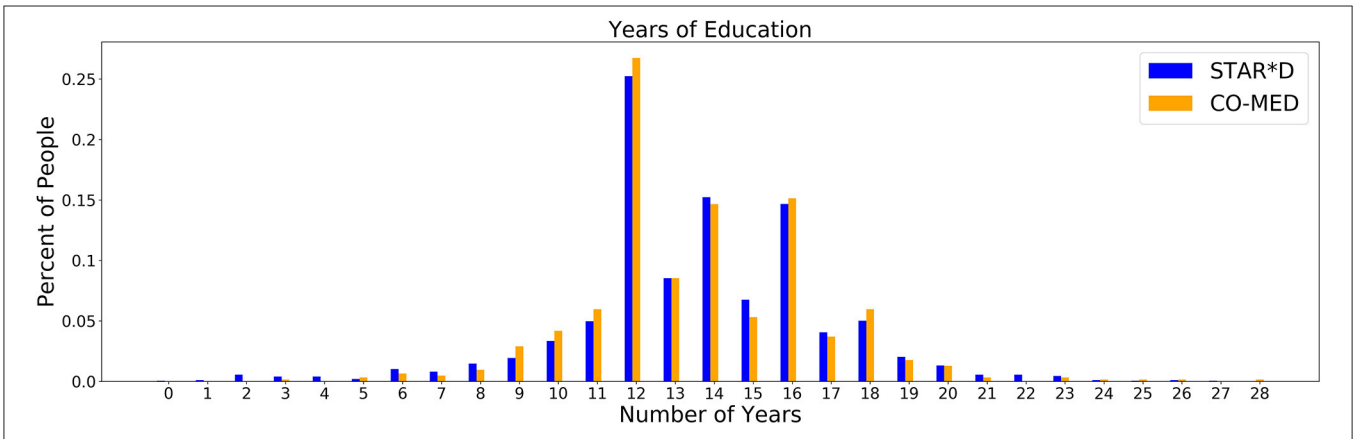


FIGURE 1 | Number of years of education for the STAR*D and CO-MED datasets.

DISCUSSION

In this manuscript we analyzed the features retained by four deep learning models of depression treatment response. We show that traditional statistics can augment the interpretation of machine learning models, while informing the nature of the underlying datasets. In addition, we offer suggestions for optimizing future data collection to improve machine-learning analyses.

Applying Insights From Machine Learning Features Toward Building Causal Mechanisms for Depression Pathology and Prognosis

Can We Identify Features Predictive of Response to Each of the Four Antidepressants Within Our Model (Escitalopram, Bupropion, Venlafaxine-Mirtazapine, Citalopram) Individually, as Well as to the Subgroup of Patients With a Low Probability of Responding to Any of the Drugs?

Across all four antidepressant subgroups, somatic energy was one of the most frequently observed features found to be in the top five features for each subject of that subgroup, consistent with previous machine learning approaches to predict response to antidepressant treatment (Chekroud et al., 2016). This may suggest that escitalopram, bupropion-escitalopram, venlafaxine-mirtazapine, and citalopram help alleviate energy symptoms (fatigue, heaviness in the body) more effectively than other symptoms. Indeed, a return of energy is often clinically observed early in treatment, and a similar effect can be observed for sad mood. Sleep-onset insomnia was also a strong predictor of response to escitalopram, bupropion-escitalopram, venlafaxine-mirtazapine, but not citalopram, suggesting that these antidepressants show some benefit in treating insomnia. However, insomnia has previously been associated with poorer treatment outcomes in some antidepressant trials (Sung et al., 2015), complicating our finding. Sleep interacts with stress to impact brain-derived neurotrophic factor levels (Giese et al., 2013), which are affected by certain antidepressants, and is also associated with other risk factors for depression, highlighting the complex interactions between depression symptomatology, risk factors like sleep, and the action of specific antidepressants. Household income was higher in the feature list of responders to each of the four antidepressant subgroups compared to the non-remission subgroup, suggesting that household income helps determine an individual's remission to any drug. This could reflect that lower income acts as a difficult-to-modify psychosocial stressor.

It should be noted that, between antidepressant categories, there were few striking differences in the symptoms more predictive of response to one treatment over another. This is consistent with the finding from the CO-MED study that there was equal efficacy of all three treatment arms. However, the model used in this analysis, detailed in Mehltretter et al. (2019), *did* find that differential treatment selection based on these features would be expected to improve population remission rates. That is, the study found that using a model trained on these features could usefully assign patients to different treatments, in a manner that suggests these treatments are not equally effective for all patients. This may be because of complex interactions between different *levels* of the different features. We may not be able to recover simple patient subtypes with the methods employed thus far. Instead, it may be the case that the subtypes that do exist include complex associations between multiple features. As a speculative example, the severity of sad mood and anxious symptoms, when combined with somatic symptoms, may have some value in determining which treatment may be most effective, over and above an analysis of the symptoms individually. We did not explore this here, but will address this question in future work. Another possibility for the lack of considerable differences in features reported in the different treatment subgroups (**Tables 4**–**8**) was the overall low number of features selected by the model. Though a low number of features was an efficient use of information when predicting remission, it was perhaps at the expense of losing some richness of explanation because it was mostly concerned with predicting remission with citalopram, the dominant drug class in the data.

We also identified features indicating a low probability of response to any of the drugs. Across all subjects with a low probability of response, initial depression severity most frequently emerged as the strongest predictor of non-response. This is consistent with extant research demonstrating increased depression severity is associated with non-response and treatment resistance (Berlim et al., 2008; De Carlo et al., 2016; Kautzky et al., 2017; Perlman et al., 2019). It suggests that the more severe the depression, the harder it will be to treat, regardless of the antidepressant. Number years education emerged as a drug-agnostic predictor of non-response. Considering its association with lack of remission (Perlman et al., 2019), low education appears an important psychosocial stressor that maintains depression, perhaps reflecting that, like low income, it is difficult to modify and therefore remains an ongoing factor that keeps people depressed for longer. Being bothered by aches and pains was also a general predictor of non-response, converging with current research on the alteration of somatic and interoceptive signaling in depression (Harshaw, 2015).

The identification of predictors of drug-specific response and general predictors of non-response to all types of treatment holds high clinical utility. Knowledge of which patients are unlikely to respond to any medication, and which will respond differentially to available first-line options will improve the treatment decision process. For instance, patients unlikely to respond to an antidepressant may consider adjunct psychotherapy, electroconvulsive therapy or intensive Day Hospital treatment earlier on in treatment, reducing prolonged symptoms of depression from ineffective treatments, potential side effects from medication, and wasted resources.

Are Specific Aspects of Trauma Predictive of Baseline Depression?

The regression analyses assessing the contribution of trauma measures to baseline depressive symptomatology found that

trauma accounts for a significant proportion of the variance in baseline depression scores, with shakiness, sweating, or heart racing from trauma reminders, an indicator of a current physical reaction related to a past trauma, presenting as a stronger contributor to baseline depression than other trauma indices, such as ever having witnessed a trauma. This indicates that while experiencing trauma does confer some vulnerability, it is those who continue to manifest symptoms—those who may have some biological or other vulnerability to the prolonged effects of trauma—who have the most depressive symptoms, and therefore a lower chance to respond to treatment. Indeed, depression is highly comorbid with post-traumatic stress disorder (PTSD) (Flory and Yehuda, 2015), suggesting a trauma-related phenotype. However, neither STAR*D nor CO-MED excluded patients with PTSD, posing the limitation that our results might be driven by patients with concomitant PTSD. Further work is needed to explore our findings and potential links with the stress-diathesis model of depression (Monroe and Simons, 1991; Colodro-Conde et al., 2018). Gender was significantly associated with baseline depression severity, consistent with higher rates of depression in females, as was the number of years of education, suggesting that low education may be an important psychosocial stressor that contributes to and, as seen in the treatment resistance modeling of question (1) above, perpetuates depression.

Do Participants in STAR*D and CO-MED Come From the Same Population, or Are These Populations Different in Key Variables That Are Predictive of Outcomes?

Education was a significant predictor in the STAR*D trial, but not in COMED. We therefore assessed whether education levels differed between the datasets, but found no significant difference. Since education is unlikely to be a drug-specific predictor, we propose that even datasets that have broad inclusion criteria and that are traditionally considered "big data" by psychiatric standards, might not be large or diverse enough to capture all of the relationships of interest between sociodemographic variables and treatment outcome.

Do Somatic Symptoms of Depression Differ by Gender?

Our analysis of somatic symptoms showed that in comparison to males, females had lower somatic energy, were more bothered by aches and pains, and had increased fatigability. This reflects current research hypothesizing that gender differences in the prevalence of depression are due to increased somatic depression among females (Silverstein et al., 2013, 2017). This points toward not only the existence of specific subtypes of depression, but also toward testable hypotheses of mechanisms for such subtypes, such as increased susceptibility to inflammation in women (Derry et al., 2015). Our results equally converge with research on the hypothalamo-pituitary-adrenal (HPA)-axis response explaining the association between stress (trauma), pain (i.e., somatic symptoms), and fatigue (McEwen, 2007).

Capturing Heterogeneity in Psychiatric Disorders: The Shift Toward "Big Diversity" in Patient Population Characteristics

Diagnostic entities in psychiatry are heterogeneous in nature, encompassing opposite ends of symptom dimensions. For major depressive disorder (MDD), diagnostic criteria can include weight gain or weight loss, increase or decrease in appetite, insomnia or hypersomnia, and psychomotor agitation or retardation (American Psychiatric Association, 2013). With 227 possible symptom combinations to meet a diagnosis of MDD (Zimmerman et al., 2015), two patients diagnosed with MDD may share no overlapping symptoms. This heterogeneity restricts the usefulness of psychiatric diagnoses for researching their etiology or prognosis, as different subtypes within a disorder might have different biological underpinnings and benefit from different types of treatment. Heterogeneity has not only hindered research, but may contribute to limited replication success in clinical trials (Dwyer et al., 2018). Traditional attempts to minimize or decompose heterogeneity include restricting inclusion criteria to focus on particular subgroups of patients (i.e., melancholic depression, treatment resistant depression, adolescent, or geriatric depression), either by imposing constraints on symptoms or limiting comorbidities, age, severity or chronicity of illness, in order to get obtain a "pure" or ideal sample of a certain subgroup to evaluate a priori hypotheses about that group. The problem with this approach is that it has not produced consistent subgroups (Marquand et al., 2016), the results may not generalize to independent samples, and such "ideal" patients are not representative of real-world heterogeneity. More optimal strategies for tackling heterogeneity may instead be data-driven approaches that capitalize on maximal heterogeneity in order to enhance generalizability of the model's predictions and mitigate bias. "Big data" requires not only large sample sizes, but "big diversity" in its samples, including multiple levels of data for each participant and variance in and across each type of data collected. Increasing data diversity will improve the generalizability and translatability of models and ensure that clinical decision aid tools might be more applicable to a broader range of individuals. Contrary to traditional approaches to experimental design in clinical populations, future research should explicitly capture variability, by including multiple study sites, ethnicities, socioeconomic levels, age, among others, to capture real-world variability and produce an ideal dataset for ML. This approach has been echoed by others and elaborated in the context of autism (Lombardo et al., 2019), but extends to all domains of mental health research. An important outcome of our deep learning model was that similar, but not identical feature sets were produced based on the sample used for training (STAR*D or COMED). For example, education, which is unlikely to be a treatment-specific predictor of response, was present in the STAR*D-dominated models, but not in the model that predicted remission in CO-MED alone, despite the average education level and the distribution of educational attainment not being significantly different between the two studies. While STAR*D was significantly larger than CO-MED, both of these datasets are considered to be large by

psychiatric research standards. The fact that one of the most key features for predicting treatment response in one dataset was not predictive of treatment response in the other provides empirical support for advocating for larger and more diverse datasets. To optimize patient outcomes with precision psychiatry, the advent of "big data" necessitates a new focus on data with "big diversity." The complexities of such data may be leveraged with ML approaches, and reinvestigated and understood with simpler, more interpretable models.

Our analyses exemplify how interpreting ML features can generate new hypotheses about disease pathology, contribute toward existing hypotheses, and help elucidate causal models which may have value in the development of new treatments or in treatment selection. Other efforts using a similar approach have proved equally fruitful: A recent study using a convolutional neural network to extract and quantify the relationship between features of the built environment and obesity prevalence showed that features of the built environment (i.e., greenery, different housing types, neighborhood density) were able to explain 64.8% of variation in obesity prevalence (Maharana and Nsoesie, 2018), demonstrating the utility of machine learning toward unpacking the association between the built environment and obesity prevalence. Through modeling complex interactions in "big data" samples, machine learning can uncover features associated with disease that can advance our understanding of psychiatric illnesses.

CONCLUSION

The analytical power of machine learning is accompanied by limitations in its interpretability. In this paper we demonstrate the benefit of using traditional statistics to improve *post-hoc* interpretation of the features selected by deep learning models trained to predict remission in depression, and can provide a more meaningful clinical interpretation to understand interrelationships between important patient demographic and clinical characteristics and depression pathology. These approaches should be viewed as hypothesis generating and not confirmatory, as the "statistical significance" (*p*-values) associated with analyses performed on variables selected via ML or indeed any variable selection approach do not retain the standard interpretation. We emphasize the advantages of investing in "big diversity"—creating large and heterogeneous datasets, instead of the homogenous datasets favored by traditional large clinical studies—in order to produce datasets that are maximally useful for addressing important clinical questions.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found using the NIMH data request and the following study identification numbers: NCT00590863; NCT00021528.

The Vulcan platform used for this work was open source and can be found here: (https://github.com/Aifred-Health/Vulcan). The data are available through the NIMH data request platform. Conflict of Interest Note: Note that we do not reproduce the final model in this work. This is because deep learning networks are difficult to represent in their final trained format, but also because the final model configuration is a trade secret of Aifred Health. However, since both the data and the platform used to train the model are open source, investigators should be able to produce versions of this model for further research purposes.

AUTHOR CONTRIBUTIONS

JM, DB, and RF contributed to the conception of the study and development of the deep learning models. JM and CR performed the statistical analyses. CR and JM wrote the first draft of the manuscript, with sections contributed by MW and DB. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was partially supported by an ERA PERMED Vision 2020 grant for the IMADAPT project. While the authors received no funding specific to this work, JM, SI, and CR were paid as consultants or employees of Aifred Health at certain points while working on this article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2019.00031/full#supplementary-material

REFERENCES

Berlim, M. T., Fleck, M. P., and Turecki, G. (2008). Current trends in the assessment and somatic treatment of resistant/refractory major depression: an overview. *Ann. Med.* 40, 149–159. doi: 10.1080/07853890701769728

Bohman, H., Jonsson, U., Paaren, A., von Knorring, L., Olsson, G., and von Knorring, A. L. (2012). Prognostic significance of functional somatic symptoms in adolescence: a 15-year community-based follow-up study of adolescents with depression compared with healthy peers. *BMC Psychiatry* 12:90. doi: 10.1186/1471-244X-12-90

Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus machine learning. *Nat. Methods* 15, 233–234. doi: 10.1038/nmeth.4642

Cao, B., Luo, Q., Fu, Y., Du, L., Qiu, T., Yang, X., et al. (2018). Predicting individual responses to the electroconvulsive therapy with hippocampal subfield volumes in major depression disorder. *Sci. Rep.* 8:5434. doi: 10.1038/s41598-018-23685-9

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., et al. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* 3, 243–250. doi: 10.1016/S2215-0366(15)00471-X

Colodro-Conde, L., Couvy-Duchesne, B., Zhu, G., Coventry, W. L., Byrne, E. M., Gordon, S., et al. (2018). A direct test of the diathesis-stress model for depression. *Mol. Psychiatry* 23, 1590–1596. doi: 10.1038/mp.2017.130

De Carlo, V., Calati, R., and Serretti, A. (2016). Socio-demographic and clinical predictors of non-response/non-remission in treatment resistant depressed patients: a systematic review. *Psychiatry Res.* 240, 421–430. doi: 10.1016/j.psychres.2016.04.034

Derry, H. M., Padin, A. C.,Kuo, J. L., Hughes, S., and Kiecolt-Glaser, J. K. (2015). Sex differences in depression: does inflammation play a role? *Curr. Psychiatry Rep.* 17:78. doi: 10.1007/s11920-015-0618-5

Dwyer, D. B., Falkai, P., and Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annu. Rev.*

Clin. Psychol. 14, 91–118. doi: 10.1146/annurev-clinpsy-032816-045037

Flory, J. D., and Yehuda, R. (2015). Comorbidity between post-traumatic stress disorder and major depressive disorder: alternative explanations and treatment considerations. *Dialogues Clin. Neurosci.* 17, 141–150.

Friedman, E. S., Wisniewski, S. R., Gilmer, W., Nierenberg, A. A., Rush, A. J., Fava, M., et al. (2009). Sociodemographic, clinical, and treatment characteristics associated with worsened depression during treatment with citalopram: results of the NIMH STAR(*)D trial. *Depress. Anxiety* 26, 612–621. doi: 10.1002/da.20568

Giese, M., Unternaehrer, E., Brand, S., Calabrese, P., Holsboer-Trachsler, E., and Eckert, A. (2013). The interplay of stress and sleep impacts BDNF level. *PLoS ONE* 8:e76050. doi: 10.1371/journal.pone.0076050

Harshaw, C. (2015). Interoceptive dysfunction: toward an integrated framework for understanding somatic and affective disturbance in depression. *Psychol. Bull.* 141, 311–363. doi: 10.1037/a0038101

Iniesta, R., Stahl, D., and McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol. Med.* 46, 2455–2465. doi: 10.1017/S0033291716001367

Jakubovski, E., and Bloch, M. H. (2014). Prognostic subgroups for citalopram response in the STAR*D trial. *J. Clin. Psychiatry* 75, 738–747. doi: 10.4088/JCP.13m08727

Kapfhammer, H. P. (2006). Somatic symptoms in depression. *Dialogues Clin. Neurosci.* 8, 227–239.

Kautzky, A., Baldinger-Melich, P., Kranz, G. S., Vanicek, T., Souery, D., Montgomery, S., et al. (2017). A new prediction model for evaluating treatment-resistant depression. *J. Clin. Psychiatry* 78, 215–222. doi: 10.4088/JCP.15m10381

Kautzky, A., Dold, M., Bartova, L., Spies, M., Vanicek, T., Souery, D., et al. (2018). Refining prediction in treatment-resistant depression: results of machine learning analyses in the TRD III sample. *J. Clin. Psychiatry* 79:16m11385. doi: 10.4088/JCP.16m11385

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., et al. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Mol. Psychiatry* 21, 1366–1371. doi: 10.1038/mp.2015.198

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., et al. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *J. Affect. Disord.* 241, 519–532. doi: 10.1016/j.jad.2018.08.073

Lin, E., Kuo, P. H., Liu, Y. L., Yu, Y. W., Yang, A. C., and Tsai, S. (2018). A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. *Front. Psychiatry* 9:290. doi: 10.3389/fpsyt.2018.00290

Lombardo, M. V., Lai, M. C., and Baron-Cohen, S. (2019). Big data approaches to decomposing heterogeneity across the autism spectrum. *Mol. Psychiatry* 24, 1435–1450. doi: 10.1038/s,41380-018-0321-0

Maharana, A., and Nsoesie, E. O. (2018). Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA Netw Open* 1:e181535. doi: 10.1001/jamanetworkopen.2018.1535

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., and Beckmann, C., F. (2016). Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 1, 433–447. doi: 10.1016/j.bpsc.2016.04.002

McEwen, B. S. (2007). Physiology and neurobiology of stress and adaptation: central role of the brain. *Physiol. Rev.* 87, 873–904. doi: 10.1152/physrev.00041.2006

Mehltretter, J., Fratila, R., Benrimoh, D., Kapelner, A., Perlman, K., Snook, E., et al. (2019). Differential treatment benefit prediction for treatment selection in depression: a deep learning analysis of STAR∗D and CO-MED data. *BioRxiv [Preprint]*. doi: 10.1101/679779

Monroe, S. M., and Simons, A. D. (1991). Diathesis-stress theories in the context of life stress research: implications for the depressive disorders. *Psychol. Bull.* 110, 406–425. doi: 10.1037//0033-2909.110.3.406

Nelson, J., Klumparendt, A., Doebler, P., and Ehring, T. (2017). Childhood maltreatment and characteristics of adult depression: meta-analysis. *Br. J. Psychiatry* 210, 96–104. doi: 10.1192/bjp.bp.115.180752

Nie, Z., Vairavan, S., Narayan, V. A.,Ye, J., and Li, Q. S. (2018). Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. *PLoS ONE* 13:e0197268. doi: 10.1371/journal.pone.0197268

Papakostas, G. I., Petersen, T., Pava, J., Masson, E., Worthington, J. J. III., Alpert, J. E., et al. (2003). Hopelessness and suicidal ideation in outpatients with treatment-resistant depression: prevalence and impact on treatment outcome. *J. Nerv. Ment. Dis.* 191, 444–449. doi: 10.1097/01.NMD.0000081591.46444.97

Passos, I. C., Mwangi, B., and Kapczinski, F. (2016). Big data analytics and machine learning: 2015 and beyond. *Lancet Psychiatry* 3, 13–15. doi: 10.1016/S2215-0366(15)00549-0

Perlis, R. H. (2013). A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biol. Psychiatry* 74, 7–14. doi: 10.1016/j.biopsych.2012.12.007

Perlman, K., Benrimoh D., Israel, S., Rollins, C., Brown, E., Tunteng, J. F., et al. (2019). A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. *J. Affect. Disord.* 243, 503–515. doi: 10.1016/j.jad.2018.09.067

Perna, G., Grassi, M., Caldirola, D., and Nemeroff, C. B. (2018). The revolution of personalized psychiatry: will technology make it happen sooner? *Psychol. Med.* 48, 705–713. doi: 10.1017/S0033291717002859

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., et al. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *Am. J. Psychiatry* 163, 1905–1917. doi: 10.1176/ajp.2006.163.11.1905

Silverstein, B., Ajdacic-Gross, V., Rossler, W., and Angst, J. (2017). The gender difference in depressive prevalence is due to high prevalence of somatic depression among women who do not have depressed relatives. *J. Affect. Disord.* 210, 269–272. doi: 10.1016/j.jad.2017.01.006

Silverstein, B., Edwards, T., Gamma, A., Ajdacic-Gross, V., Rossler, W., and Angst, J. (2013). The role played by depression associated with somatic symptomatology in accounting for the gender difference in the prevalence of depression. *Soc. Psychiatry Psychiatr. Epidemiol.* 48, 257–263. doi: 10.1007/s00127-012-0540-7

Sung, S. C., Wisniewski, S. R.,Luther, J. F.,Trivedi, M. H.,Rush, A. J., and Comed Study Team (2015). Pre-treatment insomnia as a predictor of single and combination antidepressant outcomes: a CO-MED report. *J. Affect. Disord.* 174, 157–164. doi: 10.1016/j.jad.2014.11.026

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.* 14, 779–804. doi: 10.3758/BF03194105

Williams, L. M., Rush, A. J., Koslow, S. H., Wisniewski, S. R., Cooper, N. J., Nemeroff, C. B., et al. (2011). International study to predict optimized treatment for depression (iSPOT-D), a randomized clinical trial: rationale and protocol. *Trials* 12:4. doi: 10.1186/1745-6215-12-4

Zhang, Z., Beck, M. W.,Winkler, D. A.,Huang, B., Sibanda, W., Goyal, H., et al. (2018). Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann. Transl. Med.* 6:216. doi: 10.21037/atm.2018.05.32

Zimmerman, M., Ellison, W., Young, D., Chelminski, I., and Dalrymple, K. (2015). How many different ways do patients meet the diagnostic criteria for major. *Compr. Psychiatry* 56, 29–34. doi: 10.1016/j.comppsych.2014.09.007

Conflict of Interest: DB, KP, SI, RF, and MM are shareholders of Aifred Health, a medical technology company that uses deep learning to increase treatment efficacy in psychiatry. JM, MW, and CR have received consulting fees from Aifred Health.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

HYPOTHESIS AND THEORY
published: 18 February 2020
doi: 10.3389/frai.2020.00002

Retrospective Inference as a Form of Bounded Rationality, and Its Beneficial Influence on Learning

Thomas H. B. FitzGerald[1,2,3]*, *Will D. Penny*[1,2], *Heidi M. Bonnici*[1] and *Rick A. Adams*[2,3,4]

[1] School of Psychology, University of East Anglia, Norwich, United Kingdom, [2] The Wellcome Trust Centre for Neuroimaging, University College London, London, United Kingdom, [3] Max Planck-UCL Centre for Computational Psychiatry and Ageing Research, London, United Kingdom, [4] Department of Computer Science, University College London, London, United Kingdom

Probabilistic models of cognition typically assume that agents make inferences about current states by combining new sensory information with fixed beliefs about the past, an approach known as Bayesian filtering. This is computationally parsimonious, but, in general, leads to suboptimal beliefs about past states, since it ignores the fact that new observations typically contain information about the past as well as the present. This is disadvantageous both because knowledge of past states may be intrinsically valuable, and because it impairs learning about fixed or slowly changing parameters of the environment. For these reasons, in offline data analysis it is usual to infer on every set of states using the entire time series of observations, an approach known as (fixed-interval) Bayesian smoothing. Unfortunately, however, this is impractical for real agents, since it requires the maintenance and updating of beliefs about an ever-growing set of states. We propose an intermediate approach, finite retrospective inference (FRI), in which agents perform update beliefs about a limited number of past states (Formally, this represents online fixed-lag smoothing with a sliding window). This can be seen as a form of bounded rationality in which agents seek to optimize the accuracy of their beliefs subject to computational and other resource costs. We show through simulation that this approach has the capacity to significantly increase the accuracy of both inference and learning, using a simple variational scheme applied to both randomly generated Hidden Markov models (HMMs), and a specific application of the HMM, in the form of the widely used probabilistic reversal task. Our proposal thus constitutes a theoretical contribution to normative accounts of bounded rationality, which makes testable empirical predictions that can be explored in future work.

Keywords: bayesian inference, learning, cognition, retrospective inference, reversal learning, bounded rationality, hidden markov model

OPEN ACCESS

Edited by:
Dimitrije Marković, Dresden University of Technology, Germany

Reviewed by:
Maria Chan, Memorial Sloan Kettering Cancer Center, United States
Dirk Ostwald, Freie Universität Berlin, Germany

***Correspondence:**
Thomas H. B. FitzGerald
t.fitzgerald@uea.ac.uk

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *12 July 2019*
Accepted: *14 January 2020*
Published: *18 February 2020*

Citation:
FitzGerald THB, Penny WD, Bonnici HM and Adams RA (2020) Retrospective Inference as a Form of Bounded Rationality, and Its Beneficial Influence on Learning. Front. Artif. Intell. 3:2. doi: 10.3389/frai.2020.00002

INTRODUCTION

To behave adaptively, agents need to continuously update their beliefs about present states of the world using both existing knowledge and incoming sensory information, a process that can be formalized according to the principles of probabilistic inference (von Helmholtz, 1867; Gregory, 1980). This simple insight has generated a large field of inquiry than spans most areas of the mind

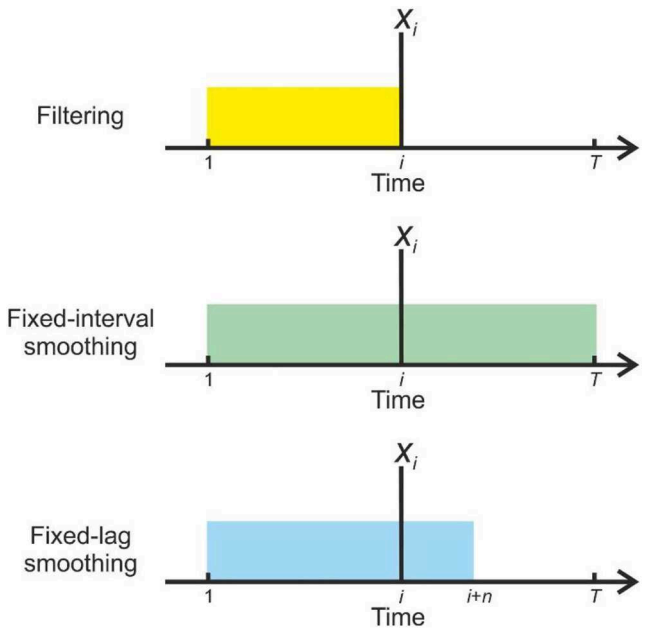


FIGURE 1 | Illustration of the information used by different strategies for inference when forming beliefs about state *x* at time *i*(indicated by the vertical line). In filtering **(Top)** beliefs are based solely on observations made up to and including that time ($\mathbf{o}_{1:i}$), as indicated by the yellow block, and are not revised in the light of subsequent information (If we assume online inference, then the present time $t = i$). In fixed-interval smoothing **(Middle)**, observations from the whole data set ($\mathbf{o}_{1:T}$) are used to inform each set of beliefs, as indicated by the green block (Here either $t = T$ or, equivalently, inference is performed offline). In fixed-lag smoothing **(Bottom)**, beliefs are retrospectively updated up to some fixed lag *n*, so $\mathbf{o}_{1:i+n}$ are used (indicated by the blue block) (See Methods for a more formal description, and explanation of the notation) (In this case $t = i + n$). Fixed-lag smoothing allows an agent to perform finite retrospective inference, which constitutes a principled trade-off between the reduced inferential accuracy resulting from filtering and the potentially severe computational costs of retrospection to an indefinite temporal depth.

and brain sciences and seeks to build probabilistic accounts of cognition (Rao and Ballard, 1999; Friston, 2010; Tenenbaum et al., 2011; Clark, 2012; Pouget et al., 2013; Aitchison and Lengyel, 2016).

In this paper, we take this framework for granted, and consider an important and related problem, that of using new sensory information to update beliefs about the past. This is important because, under conditions of uncertainty, new observations can contain significant information about past states as well as present ones (Corlett et al., 2004; Shimojo, 2014; FitzGerald et al., 2017; Moran et al., 2019).

In offline cognition or data analysis (in which agents are dealing with complete data sets, and are not required to respond to them in real time), it is possible to make inferences about all time points simultaneously (**Figure 1**).

In other words, one uses every observation to inform every belief about hidden states. This option is unavailable to real, embodied agents because they need to perceive and act in time (online) (Throughout this paper, we will denote the present time with *t*). They thus need to perform retrospective inference to increase the accuracy of their beliefs about the past. To perform retrospective inference optimally (or, equivalently in this context, to be strictly rational) it is necessary for an agent to update beliefs about a sequence of states stretching backwards to the beginning of the current task or context, or perhaps even to the beginning of its existence. This sequence is both indefinitely long and constantly growing, and representing and updating these beliefs will thus, in many situations, place intolerable demands on any real organism.

We propose an alternative approach, finite retrospective inference (FRI), in which agents update beliefs about states falling within a limited temporal window stretching into the past (FitzGerald et al., 2017). Selecting the size of this window, and thus the depth of retrospective belief updating constitutes a form of bounded rationality (Simon, 1972; Gigerenzer and Goldstein, 1996; Ortega et al., 2015), since it trades off inferential accuracy against resource costs (e.g., the metabolic and neuronal costs associated with representing beliefs, and the time to perform the calculations). The depth of updating performed by an agent in a particular context might be selected using a form of "metareasoning" in response to environmental demands (Russell and Wefald, 1991; Lieder and Griffiths, 2017). In particular, it is likely that where observations are noisier, and/or temporal dependencies are greater (in other words, where the past remains significant for longer) such strategies will be more advantageous, and are likely to be favored, provided that other constraints allow it. Alternatively, the degree of retrospection might be phenotypically specified (and thus, presumably, selected for during species evolution). In either case, a bounded-rational approach to retrospection has the potential to explain and quantify how humans and other organisms approach but do not attain optimal performance on a number of cognitive tasks.

In addition to its appeal on purely computational grounds, this proposal might help to explain the widespread occurrence of "postdictive" phenomena in perception (Eagleman and Sejnowski, 2000; Shimojo, 2014). A number of such phenomena have been noted, but in all of them perception of an event is influenced by things that only occur afterwards, suggesting a purely retrospective inference on perception (Rao et al., 2001) (Retrospective inference has also been described in the context of associative learning paradigms; Corlett et al., 2004; Moran et al., 2019). There thus seems good reason to believe that a neurobiologically plausible scheme for retrospective inference like FRI may provide valuable insights into real cognitive processes.

FRI differs from existing probabilistic accounts of online cognition (Rao and Ballard, 1999; Ma et al., 2006; Friston and Kiebel, 2009; Glaze et al., 2015; Aitchison and Lengyel, 2016), which typically only consider inferences about present states, an approach known as "Bayesian filtering" (though see Rao et al., 2001; Baker et al., 2017; Friston et al., 2017; Kaplan and Friston, 2018). It thus constitutes a novel hypothesis about cognitive function that extends probabilistic models to subsume a broader range of problems. Importantly, as we will illustrate in the simulations described below, FRI makes testable predictions about behavior and brain activity in real agents that can be tested in future experimental studies.

MATERIALS AND METHODS

Approximating Normative Inference

Consider the situation in which an agent seeks to infer on a series of T time-varying hidden states $\mathbf{x}_{1:T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ given a set of time-invariant parameters θ that are known with certainty, a series of observations $\mathbf{o}_{1:T} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$, and an initial distribution on $\mathbf{x}_0$ (Both $\mathbf{x}_{1:T}$ and $\mathbf{o}_{1:T}$ are thus random vectors). To simplify our discussion, in what follows we will assume that all the processes under consideration share the following conditional independence properties:

$$p\left(\mathbf{x}_i|\mathbf{x}_{1:i-1}, \mathbf{o}_{1:T}, \boldsymbol{\theta}\right) = p\left(\mathbf{x}_i|\mathbf{x}_{i-1}, \mathbf{o}_{i:T}, \boldsymbol{\theta}\right), \tag{1}$$

meaning that states at time i depend only upon the immediately preceding states, and are otherwise independent or previous states or observations (this is the Markov property) and that

$$p\left(\mathbf{o}_i|\mathbf{x}_{1:T}, \mathbf{o}_{1:i-1}, \mathbf{o}_{i+1:T}, \boldsymbol{\theta}\right) = p\left(\mathbf{o}_i|\mathbf{x}_i, \boldsymbol{\theta}\right), \tag{2}$$

meaning that observations depend only on the current states, and not previous states or observations. However, the general principles presented in this paper apply equally in cases where many, if not all, of these properties are relaxed, e.g., in processes with a higher-order temporal structure. We start with a general discussion of retrospective inference, which makes no specification about the nature of states and observations, before discussing a specific instantiation below (the HMM).

By the chain rule of probability, and making use of Equation 1, the joint conditional distribution over all states is given by:

$$p\left(\mathbf{x}_{1:T}|\mathbf{o}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}\right) = \prod_{i\,=\,1:T} p\left(\mathbf{x}_i|\mathbf{o}_{i:T}, \mathbf{x}_{i-1}, \boldsymbol{\theta}\right). \tag{3}$$

However, inferring on the joint distribution rapidly becomes computationally intractable, and is often unnecessary. Thus, instead of inferring on the joint conditional distribution we can instead infer on the marginal distributions over states at each time point. In other words, infer on the sequence of most likely states rather than the most likely sequence of states. This approach is known as fixed-interval Bayesian smoothing (Sarkka, 2013). The agent can thus be thought of as approximating the joint conditional distribution as:

$$p\left(\mathbf{x}_{1:T}|\mathbf{o}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}\right) \approx \prod_{i\,=\,1:T} p\left(\mathbf{x}_i|\mathbf{o}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}\right). \tag{4}$$

This provides a powerful approach for analyzing sequential data, and is widely used in offline data analysis. However, it presents serious practical difficulties for agents performing online inference of the kind that is mandatory for real, embodied agents. That is, where agents have to make inferences, and very likely take actions, whilst the process is unfolding. Specifically, it requires the agent to store and update an ever-growing set of beliefs about the past, resulting in a set of calculations that will rapidly overwhelm the cognitive capacities of plausible embodied agents. This means that "true" rationality, defined here as cognition that accords precisely with the principles of optimal probabilistic inference, is impossible for real agents, who must instead seek a feasible approximation.

In probabilistic models of online cognition, this is typically achieved by conditioning inference only on past and current observations, an approach known as Bayesian filtering (Sarkka, 2013). This means that agents make inferences of the form:

$$\prod_{i\,=\,1:T} p\left(\mathbf{x}_i|\mathbf{o}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}\right) \approx \prod_{i\,=\,1:T} p\left(\mathbf{x}_i|\mathbf{o}_{1:i}, \mathbf{x}_0, \boldsymbol{\theta}\right). \tag{5}$$

From the perspective of disembodied normative inference, the approximation implied here represents suboptimality. However, for a real cognitive agent, it can be thought of as an unavoidable cost of having to perform inference in time, which necessitates the use of an alternate strategy.

Filtering can be implemented in a straightforward fashion by recursive application of:

$$\begin{aligned} &p\left(\mathbf{x}_i|\mathbf{o}_{1:i}, \mathbf{x}_0, \boldsymbol{\theta}\right) \\ &= \int \frac{p\left(\mathbf{o}_i|\mathbf{x}_i, \boldsymbol{\theta}\right) p\left(\mathbf{x}_i|\mathbf{x}_{i-1}, \boldsymbol{\theta}\right)}{p\left(\mathbf{o}_i|\mathbf{x}_0, \boldsymbol{\theta}\right)} p\left(\mathbf{x}_{i-1}|\mathbf{o}_{1:i-1}, \mathbf{x}_0, \boldsymbol{\theta}\right) d\mathbf{x}_{i-1}. \end{aligned} \tag{6}$$

It is thus computationally parsimonious, since it requires only a single set of calculations at each time step and only requires an agent to store fixed beliefs about the past. In the case of first-order processes, this is only about the immediately preceding time step. However, this parsimony comes at a cost, since it reduces the accuracy of an agent's beliefs about the past, and consequently, as will be discussed later, impairs learning.

To remedy this, an agent that is performing Bayesian filtering, can implement smoothing recursively by performing an additional "backwards pass" through the data:

$$\begin{aligned} &p\left(\mathbf{x}_i|\mathbf{o}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}\right) \\ &= p\left(\mathbf{x}_i|\mathbf{o}_{1:i}, \mathbf{x}_0, \boldsymbol{\theta}\right) \int \left[\frac{p\left(\mathbf{x}_{i+1}|\mathbf{x}_i\right) p\left(\mathbf{x}_{i+1}|\mathbf{o}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}\right)}{p\left(\mathbf{x}_{i+1}|\mathbf{o}_{1:i}, \mathbf{x}_0, \boldsymbol{\theta}\right)}\right] d\mathbf{x}_{i+1}. \end{aligned} \tag{7}$$

(Use of an integral here and in Equations 8, 10 presupposes that states are continuous-valued. In the case of discrete states, as in the HMM discussed below, this is replaced with a summation). Here $p\left(\mathbf{x}_i|\mathbf{o}_{1:i}, \mathbf{x}_0, \theta\right)$ is the state estimate derived from filtering, $p\left(\mathbf{x}_{i+1}|\mathbf{x}_i\right)$ is the dynamic model governing transitions between states, $p\left(\mathbf{x}_{i+1}|\mathbf{o}_{1:T}, \mathbf{x}_0, \theta\right)$ is the smoothed state estimate at $i+1$, and $p\left(\mathbf{x}_{i+1}|\mathbf{o}_{1:i}, \mathbf{x}_0, \theta\right)$ is the predicted distribution at $i+1$ given by:

$$p\left(\mathbf{x}_{i+1}|\mathbf{o}_{1:i}, \mathbf{x}_0, \boldsymbol{\theta}\right) = \int p\left(\mathbf{x}_{i+1}|\mathbf{x}_i\right) p\left(\mathbf{x}_i|\mathbf{o}_{1:i}, \mathbf{x}_0, \boldsymbol{\theta}\right) d\mathbf{x}_i. \tag{8}$$

Thus (fixed-interval), smoothing can be carried out in a straightforward manner, beginning with the current state estimate derived from filtering, and working iteratively backwards. Nonetheless, it requires the agent to perform a set of calculations that grows linearly with the time series, and store a similarly growing set of beliefs about past states, and thus introduces significant extra costs for an agent over and above filtering, which are likely to become unsustainable for real agents in ecological contexts. We thus propose that agents make use of

an intermediate strategy, finite retrospective inference, in which they perform retrospective belief updating to a limited degree, in a manner that reflects both the desirability of accurate inference and the need to limit resource (and other) costs.

Finite Retrospective Inference

To implement FRI, we propose that agents perform fixed-lag smoothing, an approach that is intermediate between full (fixed-interval) smoothing and filtering. In fixed-lag smoothing, agents update beliefs about all states within a fixed-length time window that includes the present time but stretches a set distance into the past (**Figure 1**) (FitzGerald et al., 2017). This window moves forward in time at the same rate that observations are gathered, meaning that cognition occurs within a sliding window (In principle the sliding window approach can also be used to infer on the joint distribution of short sequences of states FitzGerald et al., 2017, but we focus on smoothing in this paper for the sake of simplicity). We are unaware of a precedent for this approach in treatments of cognition, however it has been employed in other contexts (Moore, 1973; Cohn et al., 1994; Chen and Tugnait, 2001; Sarkka, 2013). This means that, for a window of length considered at time t, agents approximate the true marginal distribution as follows:

$$\prod_{i\,=\,1:T} p\left(\mathbf{x}_i|\mathbf{o}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}\right) \approx \prod_{i\,=\,1:T} p\left(\mathbf{x}_i|\mathbf{o}_{1\,:\,i+n-1}, \mathbf{x}_0, \boldsymbol{\theta}\right). \tag{9}$$

As can be seen by comparing Equations (5, 9), filtering is a special case of fixed-lag smoothing in which $n = 1$. Smoothing can thus be performed by iteratively evaluating.

$$\prod_{i\,=\,t-n+1}^{t} p\left(\mathbf{x}_i|\mathbf{o}_{1:t}, \mathbf{d}, \boldsymbol{\theta}\right) = \int p\left(\mathbf{x}_{t-n}|\mathbf{o}_{1:t-n}, \mathbf{d}, \boldsymbol{\theta}\right) \prod_{i\,=\,t-n+1}^{t} p\left(\mathbf{x}_i|\mathbf{o}_{t-n+1:t}, \mathbf{x}_{t-n}, \mathbf{d}, \boldsymbol{\theta}\right) d\mathbf{x}_{t-n}. \tag{10}$$

This simply requires the agent to track $p\left(\mathbf{x}_{t-n}|\mathbf{o}_{1:t-n}, \mathbf{d}, \theta\right)$, the filtered estimate of the states that obtain at the timestep immediately preceding the current window. Practically, fixed-lag smoothing can be implemented using Equation (7), with the proviso that backward recursion is only performed $n - 1$ times. In other words, rather than propagating new information right the way back through a time series as is typical in offline applications, it is only propagated to a fixed depth $(n-1)$, limiting the computational cost to the agent. This allows agents to adopt a bounded rational strategy in which they trade off inferential accuracy and computational (and potentially other) costs to select an appropriate depth of processing.

Parameter Learning Using Retrospective Inference

We next consider the more general situation in which there is uncertainty about both states and parameters, and agents must therefore perform learning as well as inference. This is often referred to as a "dual estimation" problem (Wan et al., 1999; Friston et al., 2008; Radillo et al., 2017), and is characteristic of many real-world situations. To do so, we make the model parameters θ random variables, and condition beliefs about them on a set of fixed hyperparameters λ, such that states and observations are independent of the hyperparameters when conditioned on the parameters, meaning that:

$$p\left(\mathbf{x}_{1:T}, \boldsymbol{\theta}|\mathbf{o}_{1:T}, \boldsymbol{\lambda}\right) = p\left(\mathbf{x}_{1:T}|\mathbf{o}_{1:T}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|\mathbf{o}_{1:T}, \boldsymbol{\lambda}\right). \tag{11}$$

Learning and inference are inextricably related to one another, since beliefs about states depend on beliefs about parameters, and vice versa. Since parameters are fixed, accurately estimating them involves accumulating evidence across entire time series, and thus beliefs about multiple sets of states. This means that increasing the accuracy of beliefs about the past, through retrospective belief updating, also increases the accuracy of parameter estimation. Crucially, improved parameter estimates will also result in more accurate beliefs about the present and better predictions about the future. Thus, in the context of uncertainty about model parameters, retrospective belief updating is advantageous even for an agent that has no intrinsic interest in the past. This is a very important point, since it argues for the wide importance of retrospective belief updating across a variety of situations and agents.

At a practical level, learning using FRI is very similar to offline learning. We treat each window as a time series in its own right, with λ is replaced by $\tilde{\lambda}$, the sufficient statistics of $p\left(\theta|\mathbf{o}_{1:t-n}, \lambda\right)$, which is the posterior distribution over the parameters conditioned on all observations preceding the current window, and perform learning and inference as normal. The use of a sliding window does, however introduce a small additional complexity, since successive windows overlap and thus share data points. Thus, if we treated each window as a separate time series we would count each observation multiple times and, as a result, overweight them. To avoid this, when updating $\tilde{\lambda}$ we only use information about states at the first time-point in the window (i.e., $p\left(\mathbf{x}_{t-n+1}|\mathbf{o}_{1:t}, \theta\right)$) (A more specific example of this is provided for the HMM below). This also means that only the best available estimate of each set of states (in other words, the estimate that will not be revised in light of future evidence) contributes to stored beliefs about the parameters of the model.

Retrospective Inference in Hidden Markov Models

To illustrate the utility of bounded-rational retrospective inference for an agent, we applied the principles described above to Hidden Markov models (**Figure 2**). In principle though, they apply equally to a broad range of models with alternative properties such as continuous state spaces and higher-order temporal structure. In an HMM, the system moves though a series of T time-varying hidden states, each of which is drawn from a discrete state space of dimension K. Hidden states $x_{1:T}$ are not observed directly, but instead must be inferred from observed variables. Here we assume that these also discrete, with dimension M, but this need not be the case. Thus, at time t (where $t \in \mathbb{N} : t \in \{1, T\}$), x_t is a binary vector of length K such that $\sum x_t = 1$, and similarly o_t is a binary vector of length M

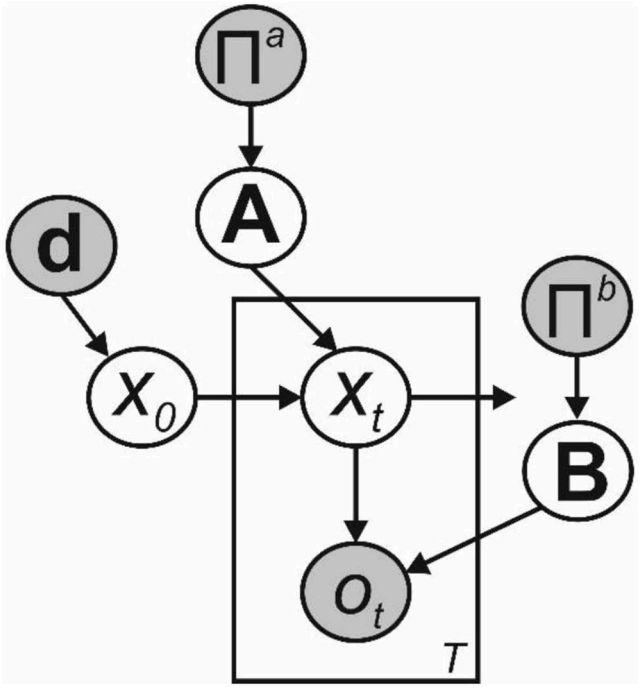


FIGURE 2 | Bayesian graph illustrating the structure of the Hidden Markov model described in the text (Shaded circles indicate variables with known values, unshaded circles indicate hidden variables). Transitions between hidden states x_0 to x_T are governed by the transition matrix **A**, and are first-order Markovian. Observations o_1 to o_T depend only on the current hidden state and the emission matrix **B**. Where the parameters of **A** and **B** need to be learnt, as depicted here we include appropriate sets of Dirichlet priors, parameterized by the matrices Π^a and Π^b, respectively. Beliefs about the initial hidden state x_0 are governed by the parameter vector **d**.

such that $\sum o_t = 1$ (This can also be described as a multinoulli random variable).

Initial state probabilities are encoded in a row vector d, which we will hereafter assume to encode a uniform distribution. Transitions between states are first-order Markovian, and the transition probabilities are encoded in a $K \times K$ matrix **A**, such that:

$$A_{jk} \equiv p\left(x_{i,k} = 1 | x_{i-1,j} = 1\right), \tag{12}$$

where, $A_{jk} \in [0, 1]$ and $\sum_k A_{jk} = 1$. This means that each row $\mathbf{A}_{j\bullet}$ encodes the transition probabilities from state j to the entire state space. From this it follows (Bishop, 2006) that:

$$p\left(x_i | x_{i-1}, \mathbf{A}\right) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{x_{i-1,j} x_{ik}}. \tag{13}$$

Similarly, the $M \times K$ matrix **B** encodes the emission probabilities such that:

$$B_{jk} \equiv p\left(o_{i,k} = 1 | x_{i,j} = 1\right), \tag{14}$$

where $B_{jk} \in [0, 1]$ and $\sum_k B_{jk=1}$. Thus, that each row $\mathbf{B}_{j\bullet}$ encodes the probability of each observed variable when in state j, and:

$$p\left(o_i | x_i, \mathbf{B}\right) = \prod_{k=1}^{M} \prod_{j=1}^{K} B_{jk}^{x_{ij} o_{ik}}. \tag{15}$$

Pure Inference in HMMs

To calculate the smoothed marginal posterior $\gamma\left(x_i\right)$ in an HMM, we can make use of the forward-backward algorithm (Rabiner, 1989). This involves recursive forward and backward sweeps, that calculates two quantities $\alpha\left(x_i\right)$ and $\beta\left(x_i\right)$ for each time point (Bishop, 2006) such that:

$$\begin{aligned} \gamma\left(x_i\right) &= p\left(x_i | o_{1:T}, \mathbf{A}, \mathbf{B}\right), \\ &= \frac{\alpha\left(x_i\right) \beta\left(x_i\right)}{\sum \alpha\left(x_i\right) \beta\left(x_i\right)}, \\ \alpha\left(x_i\right) &\equiv p\left(o_{1:i}, x_i | \mathbf{A}, \mathbf{B}\right), \\ \beta\left(x_i\right) &\equiv p\left(o_{i+1:T} | x_i, \mathbf{A}, \mathbf{B}\right). \end{aligned} \tag{16}$$

$\alpha\left(x_i\right)$ thus corresponds to the unnormalized filtered posterior, and is given by:

$$\alpha\left(x_1\right) = \left(\mathbf{B} o_1\right) \circ \mathbf{d}, \tag{17}$$

for the first state, and:

$$\alpha\left(x_i\right) = \left(\mathbf{B} o_i\right) \circ \left(\mathbf{A}^{\mathrm{T}} \alpha\left(x_{i-1}\right)\right), \tag{18}$$

for all subsequent states. Here $\circ$ denotes the Hadamard or element-wise product. $\beta\left(x_i\right)$ is given by:

$$\beta\left(x_i\right) = \mathbf{A}\left(\beta\left(x_{i+1}\right) \circ \left(\mathbf{B} o_{i+1}\right)\right). \tag{19}$$

To apply the sliding window approach to this model, at each timestep we simply evaluate the filtered posterior using and then perform backward inference a fixed number of steps using. This is the key step that enables the agent to perform FRI by inferring over both the present state and a sequence of previous states stretching a fixed distance into the past.

Dual Estimation in HMMs

To learn the transition probabilities of an HMM we first need to define an additional quantity, the dual-slice marginal $\xi\left(x_i, x_{i-1}\right)$, which corresponds to the joint probability distribution $p\left(x_i, x_{i-1} | o_{1:T}, \theta\right)$ (Baum et al., 1970; Bishop, 2006). It is simple to show that:

$$\xi\left(x_i, x_{i-1}\right) \propto \mathbf{A} \circ \left(\alpha\left(x_{i-1}\right)\left(\left(\mathbf{B} o_i\right) \circ \beta\left(x_i\right)\right)^{\mathrm{T}}\right). \tag{20}$$

(For a more detailed exposition of this see Bishop, 2006).

Introducing learning renders exact inference impossible, which necessitates the use of an approximation. Broadly speaking, such approximations fall into two categories: sampling approaches (Andrieu et al., 2003), which are computationally expensive but asymptotically exact, and variational approaches which are more computationally efficient but require the

introduction of a tractable approximate distribution (Blei et al., 2017). We focus here on implementing model inversion using variational Bayes (Beal, 2003), which we believe has some neurobiological plausibility (Friston et al., 2017). This is not a strong claim, however, about the actual mechanisms used by human observers (or indeed any other agent), and similar results could be derived under any appropriate scheme (see **Appendix** for further description of the variational methods employed here).

In the offline case, this model has been described in Mackay (1997) and Beal (2003), and the reader is referred to these sources for more detailed expositions. Briefly, we start by placing Dirichlet priors over each row of the transition matrix **A** and the observation matrix **B** such that:

$$\begin{aligned} p\left(\mathbf{A}_{j\bullet}\right) &= Dir\left(\Pi^{a}_{j\bullet}\right), \\ \mathbb{E}\left[a_{jk}\right] &= \frac{\pi^{a}_{jk}}{\sum_{k=1}^{K}\pi^{a}_{jk}}, \\ \mathbb{E}\left[\ln a_{jk}\right] &= \psi\left(\pi^{a}_{jk}\right) - \psi\left(\sum_{k=1}^{K}\pi^{a}_{jk}\right). \\ p\left(\mathbf{B}_{j\bullet}\right) &= Dir\left(\Pi^{b}_{j\bullet}\right), \\ \mathbb{E}\left[b_{jk}\right] &= \frac{\pi^{b}_{jk}}{\sum_{k=1}^{K}\pi^{b}_{jk}}, \\ \mathbb{E}\left[\ln b_{jk}\right] &= \psi\left(\pi^{b}_{jk}\right) - \psi\left(\sum_{k=1}^{K}\pi^{b}_{jk}\right). \end{aligned} \tag{21}$$

where Π^{a} and Π^{b} are matrices encoding the concentration parameters of the Dirichlet distributions, and ψ is the digamma function. Since the Dirichlet distribution is the conjugate prior for a multinomial likelihood, this enables us to carry out parameter learning using a set of simple update equations as described below.

The log joint probability distribution for the model thus becomes,

$$\begin{aligned} \ln p\left(o_{1:T}, x_{1:T}, \mathbf{A}, \mathbf{B}|\mathbf{d}, \mathbf{\Pi}^{a}, \mathbf{\Pi}^{b}\right) &= \sum_{i=1}^{T}\ln p\left(o_i|x_i, \mathbf{B}\right) \\ &+ \sum_{i=2}^{T}\ln p\left(x_i|x_{i-1}, \mathbf{A}\right) \\ &+ \ln p\left(\mathbf{B}|\mathbf{\Pi}^{b}\right) + \ln p\left(\mathbf{A}|\mathbf{\Pi}^{a}\right) \\ &+ \ln p\left(x_1|\mathbf{d}\right), \end{aligned} \tag{22}$$

and model inversion can be performed by iteratively evaluating the following update equations for the states and parameters (see **Appendix** for a full derivation).

$$\begin{aligned} \widehat{\boldsymbol{X}}_i &\equiv \mathbb{E}_{\mathbf{A},\mathbf{B}}\left[\gamma\left(x_i\right)\right] \\ &= \left(\left(\widehat{\boldsymbol{B}}o_i\right) \circ \left(\widehat{\boldsymbol{A}}^{\mathrm{T}}\,\alpha\left(x_{i-1}\right)\right)\right)\left(\widehat{\boldsymbol{A}}\left(\beta\left(x_{i+1}\right) \circ \left(\widehat{\boldsymbol{B}}o_{i+1}\right)\right)\right), \\ \widehat{\boldsymbol{M}}_i &\equiv \mathbb{E}_{\mathbf{A},\mathbf{B}}\left[\xi\left(x_{i-1}, x_i\right)\right] \\ &\propto \widehat{\boldsymbol{A}} \circ \left(\alpha\left(x_{i-1}\right)\left(\left(\widehat{\boldsymbol{B}}o_i\right) \circ \beta\left(x_i\right)\right)^{\mathrm{T}}\right), \\ \ln \widehat{\boldsymbol{a}}_{jk} &\equiv \psi\left(\widehat{\boldsymbol{\pi}}^{a}_{jk}\right) - \psi\left(\sum_{k=1}^{K}\widehat{\boldsymbol{\pi}}^{a}_{jk}\right) \\ \widehat{\mathbf{\Pi}}^{a} &\equiv \mathbf{\Pi}^{a} + \sum_{i=2}^{T}\widehat{\boldsymbol{M}}_i, \\ \ln \widehat{\boldsymbol{b}}_{jk} &\equiv \psi\left(\widehat{\boldsymbol{\pi}}^{b}_{jk}\right) - \psi\left(\sum_{k=1}^{K}\widehat{\boldsymbol{\pi}}^{b}_{jk}\right), \\ \widehat{\boldsymbol{\pi}}^{b} &\equiv \mathbf{\Pi}^{b} + \sum_{i=1}^{T}\widehat{\boldsymbol{\pi}}_i\mathbf{o}_i^{\mathbf{T}}. \end{aligned} \tag{23}$$

(Here the "hat" notation denotes expectations of the distributions over hidden variables generated using the variational inference scheme). This means that inference about the smoothed $\gamma\left(x_i\right)$ and dual-slice marginals $\xi\left(x_i, x_{i-1}\right)$ is calculated by applying the forward-backward algorithm at each iteration, using the variational estimates $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$, in place of the non-Bayesian **A** and **B** used in Equations (17–19) (Mackay, 1997). The update equations for the parameters also have intuitive interpretations. Updates of the transition matrix **A** correspond to accumulating evidence about the number of times each state transition occurs, whilst those for the observation matrix **B** correspond to a similar evidence accumulation process, this time about the number of times that a particular observation was made whilst occupying a particular state.

For the variational HMM, the lower bound L can be calculated in terms of the normalization constants $\sum\alpha\left(x_i\right)$ derived during filtering, and the Kullback-Leibler divergences between prior and posterior distributions over the parameters (see Beal, 2003; Bishop, 2006 for derivations). Thus,

$$\begin{aligned} L = &\sum_{i=1}^{I}\ln\left(\sum\alpha\left(x_i\right)\right) - \sum_{k=1}^{K}D_{KL}\left(Dir\left(\widehat{\mathbf{\Pi}}^{a}_{jk}\right)||Dir\left(\mathbf{\Pi}^{a}_{jk}\right)\right) \\ &- \sum_{k=1}^{K}D_{KL}\left(Dir\left(\widehat{\mathbf{\Pi}}^{b}_{jk}\right)||Dir\left(\mathbf{\Pi}^{b}_{jk}\right)\right). \end{aligned} \tag{24}$$

In all simulations, iterations were performed until the difference in the variational lower bound L was $<1^{-6}$ times the number of data points (T). To carry out online learning and inference, we simply apply the sliding window approach described earlier to this model. This means that we only evaluate $\widehat{\mathbf{x}}_j$ and $\widehat{\mathbf{M}}_j$ for timepoints that fall within the current window, and

parameter learning is performed by updating the concentration parameters using the following equations (where t indicates the present time):

$$\begin{aligned}\tilde{\mathbf{\Pi}}^a &= \mathbf{\Pi}^a + \widehat{\boldsymbol{M}}_{t-n+1},\\ \widehat{\mathbf{\Pi}}^a &= \tilde{\mathbf{\Pi}}^a + \sum_{j=t-n+2}^{t} \widehat{\mathbf{M}}_j,\\ \tilde{\mathbf{\Pi}}^b &= \mathbf{\Pi}^b + \widehat{\boldsymbol{x}}_{t-n+1}\widehat{\mathbf{o}}_{t-n+1}{}^{\mathbf{T}},\\ \widehat{\mathbf{\Pi}}^b &= \tilde{\Pi}^b + \sum_{j=t-n+2}^{t} \widehat{\boldsymbol{x}}_j\widehat{\mathbf{o}}_j{}^{\mathbf{T}}.\end{aligned} \tag{25}$$

Here $\tilde{\Pi}^a$ and $\tilde{\Pi}^b$ denote the fixed-lag parameters that are incremented across time steps, Π^a and Π^b denote the values of the fixed-lag concentration parameters from the previous time step (in other words, the evidence that has been accumulated prior to the current window), and $\widehat{\Pi}^a$ and $\widehat{\Pi}^b$ denote the full estimates of the concentration parameters based on timesteps 1 to t.

The Probabilistic Reversal Task as a Special Case of the HMM

To illustrate the utility of FRI even for relatively straightforward tasks, we simulated inference and learning on a probabilistic reversal paradigm (Hampton et al., 2006; Glaze et al., 2015; Radillo et al., 2017). Briefly, subjects are required to track an underlying hidden state that occasionally switches between one of two possible values, based on probabilistic feedback (In other words, feedback that is only, for example, 85% reliable). This paradigm is both simple and widely used, and the small state space makes illustrating results in graphical form relatively straightforward. In addition, the fact that the paradigm is widely used makes it an appealing tool for exploring to what extent human subjects actually employ FRI when solving this sort of task. The task can be modeled as an HMM, in which there are only two hidden states, which probabilistically generate one of two possible observations (Hampton et al., 2006; Schlagenhauf et al., 2014; Costa et al., 2015; FitzGerald et al., 2017). The parameter r encodes the probability of a reversal between trials, and v encodes the reliability of observations. Thus,

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} 1-r & r \\ r & 1-r \end{bmatrix},\\ \mathbf{B} &= \begin{bmatrix} v & 1-v \\ 1-v & v \end{bmatrix}.\end{aligned} \tag{26}$$

(Introducing learning requires a slight modification of the standard HMM parameter update equations to reflect the symmetry of the **A** and **B** matrices, as described in the **Appendix**).

Simulations

Probabilistic Reversal Task

To illustrate the effects of retrospective inference on an agent's beliefs whilst doing the probabilistic reversal task, we simulated 1,000 instantiations of a 256 trial task session, with parameters set as $r = 0.1$ and $v = 0.85$, plausible values for real versions of the task (e.g., FitzGerald et al., 2017). For the "pure inference" agent, we set extremely strong (and accurate) prior beliefs about **A** and **B** by setting initial values of $\Pi^{(a)} = 10^6\mathbf{A}$ and $\Pi^{(b)} = 10^6\mathbf{B}$. This has the consequence of effectively fixing these parameters to their prior values (in other words, essentially rendering them fixed parameters). For the "dual estimation" agent, we kept the prior beliefs about **B** identical, but set weak priors on the transition matrix of:

$$\Pi^{(a)} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}. \tag{27}$$

This has the consequence of allowing agents' beliefs to be determined almost completely by the data they encounter. Window lengths for retrospective inference were set at 1, 2, 4, 8, 16, 32, 64, 128, and 256 trials, and we also simulated an agent performing offline (fixed interval) smoothing for comparison. To assess the accuracy of inference and learning, we calculated the log likelihood assigned to the true sequence of hidden states and the true value of r, calculated using Equation 36 (see **Appendix**), and averaged these across simulations.

The aim of these simulations is to demonstrate the effects, and potential advantages, of performing FRI for an agent, even on relatively simple tasks. However, establishing whether retrospective inference is in fact a feature of human cognition requires careful experimental validation. This will involve careful model-based analysis of behavioral (and possibly neuroimaging) data collected on appropriate behavioral tasks. We intend to address this in future studies.

Random HMMs

To show that the effects that we illustrate are not due to some specific feature of the probabilistic reversal paradigms, we performed similar simulations, this time using HMMs with three possible hidden states, three possible observations, and randomly generated transition probabilities. We generated 10 such HMMs, and simulated 100 instantiations of each, whilst varying the diagonal terms of the emission matrix **B** at intervals of 0.05 between 0.65 and 0.95 (and setting the off-diagonal terms to be equal) (This corresponds to varying the degree of perceptual uncertainty). Prior beliefs for the pure inference and dual estimation agents were set as described for the reversal task, and accuracy was assessed in a similar manner, using Equation 23.

RESULTS

To explore the properties of fixed-lag retrospection in pure inference problems (in other words, ones where no learning is necessary), we simulated behavior on both the probabilistic reversal task and on random HMMs. As expected, in both cases, FRI considerably improved the accuracy of agents' final (offline) beliefs about past hidden states. (Online estimates of current states are identical under all approaches). Strikingly, in both cases, this improvement occurred even when agents only retrospected over short windows (**Figure 3**), suggesting that, in

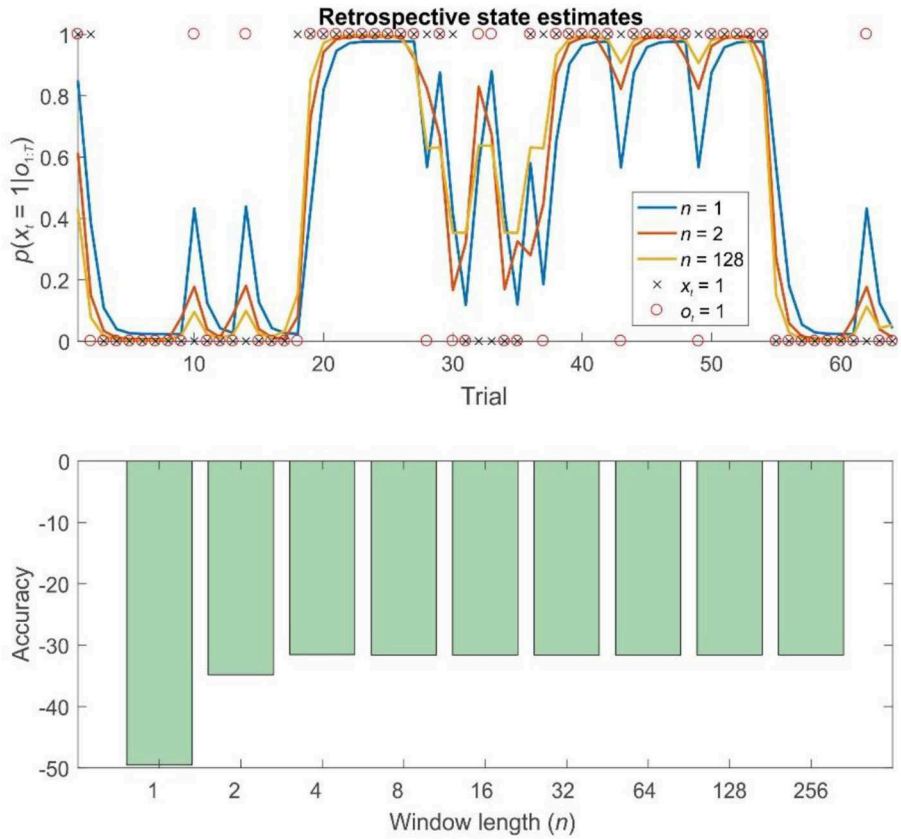


FIGURE 3 | Retrospective belief updating improves state estimates during pure inference on a probabilistic reversal task. Top panel: illustration of the first 64 trials of a 256 trial session of the reversal task using different strategies. The final (retrospective) posteriors are shown in blue ($n = 1$, filtering), orange ($n = 2$), and gold ($n = 128$). Black crosses show the true hidden state, and red circles the observations made on each trial. Retrospective belief updating allows agents to infer the true underlying states more accurately. Bottom panel: relative log accuracy of models of different window lengths, averaged across simulated time series (see main text for details) (Accuracy is quantified as the log likelihood assigned to the true sequence of states by the agent, averaged across simulations). This illustrates that, in this context at least, even the use of a very short window leads to significantly more accurate beliefs, but that this benefit saturates relatively rapidly (by about $n = 8$). Thus, for a bounded rational agent performing pure inference, the optimal window length may be surprisingly low, depending on the relevant computational costs.

certain problems at least, a limited capacity for retrospection can yield significantly improves inference.

Simulations of dual estimation problems in which there was uncertainty about r clearly illustrated that retrospective inference increases the accuracy of both retrospective *and* online state estimation, as a result of increased accuracy in parameter learning (**Figures 4**, **5**). One important feature to note is that even when the maximum possible depth of retrospection is employed ($n = 256$), the accuracy of online state estimation always falls significantly short of offline estimation. This indicates the fact that, however great the representational and computational sophistication of an agent, there is always a cost to performing inference online, rather than with a complete data set. If sufficiently high, this cost provides an incentive to perform additional (subsequent) offline processing, perhaps during sleep, and it is conceivable that this might be linked to the extended process of memory consolidation. Similar patterns were observed in the random HMM simulations, supporting the notion that these are general properties of retrospective inference (**Figure 6**).

DISCUSSION

In this paper, we consider the problem of accurately updating beliefs about the past from the perspective of probabilistic cognition. Specifically, we propose that humans and other agents use finite retrospective inference, in which beliefs about past states are modifiable within a certain temporal window, but are fixed thereafter. We show, using simulations of inference and learning in the context of a probabilistic reversal task, that even a fairly limited degree of retrospection results in significantly improved accuracy of beliefs about both states and parameters. Importantly, the hypothesis that agents perform retrospective inference makes clear predictions about behavior on appropriate tasks that are quantitatively dissociable from those made under the hypothesis that agents use pure filtering. Implementing retrospective inference also makes specific predictions about brain function, since it requires beliefs about past states to be explicitly represented and updated. Our work thus provides testable hypotheses that can be explored in future behavioral and neurobiological studies.

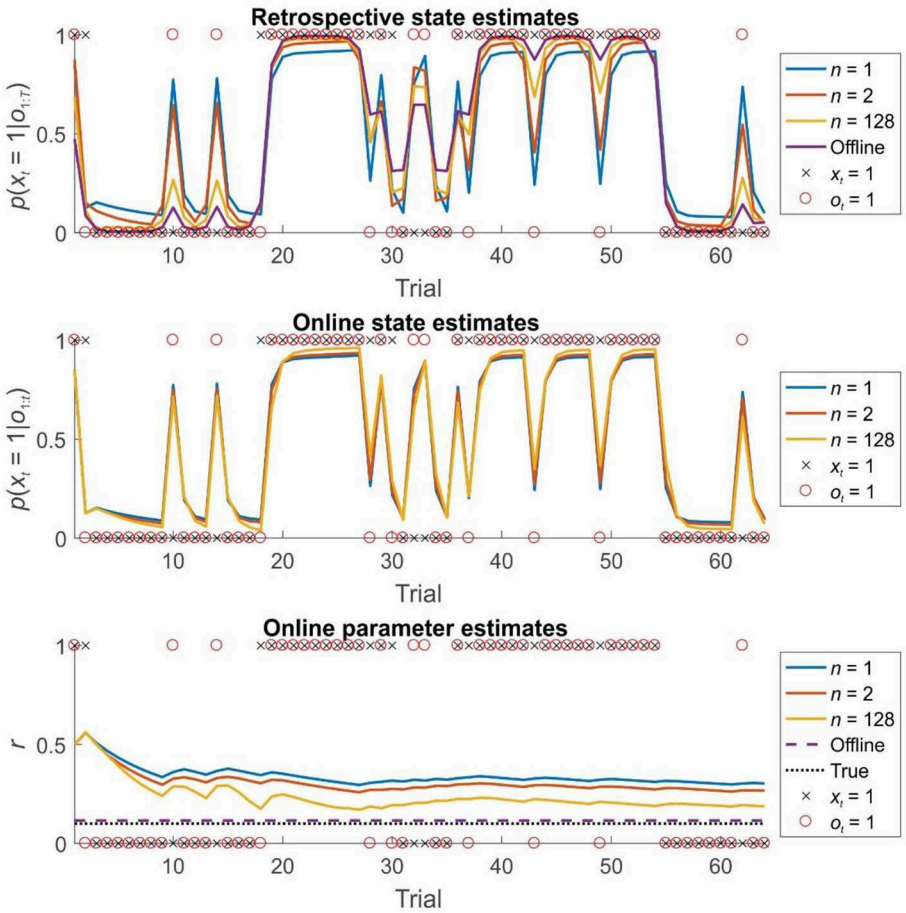


FIGURE 4 | State and parameter estimation for agents performing dual estimation on the first 64 trials of a 256 trial session of the reversal task using different strategies. Top panel: the accuracy of retrospective belief estimates $p(x_i|o_{1:T})$ increases with greater window lengths, but still falls well-short of the performance of an offline agents, who has access to the entire time series simultaneously. Middle panel: the accuracy of online (filtered) beliefs about the current state $p(x_i|o_{1:i})$ subtly but consistently increases with greater window length. Note that this effect is entirely due to the beneficial effects of greater window lengths on parameter learning. Bottom panel: the effect of window length on parameter learning. Estimates of r are derived from Π^a at each timestep (the best estimate available to the agent at that time). With greater window lengths, parameter estimates converge more rapidly on the true value (Estimates from agents performing retrospective inference with windows of length 1, 2, and 128 time steps are shown in blue, orange, and gold, respectively. Estimates from an agent performing offline inference is shown in purple. True hidden states are indicated with black crosses, whilst observations are indicated with red circles. The true value of parameter r is indicated with a dotted black line in 5c).

Perhaps the most significant feature of our simulations is the demonstration that, where there is uncertainty about time-invariant model parameters, finite retrospective inference significantly improves the accuracy of learning. This is important both because these parameters may be of intrinsic interest, and because better learning will result in more accurate beliefs about present and future states. Even if an agent has no intrinsic interest in past events, it still has a clear incentive to perform retrospective inference, since this will allow it to act better in the future. This provides a new twist on the often-advanced hypothesis that the primary function of memory in general, and episodic memory in particular, is to improve predictions about the future (Schacter et al., 2012). Here, in addition to playing a role in constructing imagined future scenarios (Hassabis et al., 2007), the explicit representation of events or episodes in the past may be essential for updating beliefs about current and future states or learning time-invariant properties of an agent's environment (Baker et al., 2017). In this paper we have not sought to clearly characterize the sorts of problem for which FRI is likely to be most useful. However, this will be extremely important for future work aimed at furnishing empirical evidence for an effect of retrospective inference on learning.

A similar point may be made about the potential importance of retrospective inference for the generation and selection of appropriate cognitive models, a process known as structure learning (Acuña and Schrater, 2010; Braun et al., 2010; Tervo et al., 2016). In this paper, we confine ourselves to considering inference about hidden states and learning about fixed model parameters, but structure learning is an equally important process, and one that is likely to be strongly affected by the depth of retrospective processing employed by an individual. In future

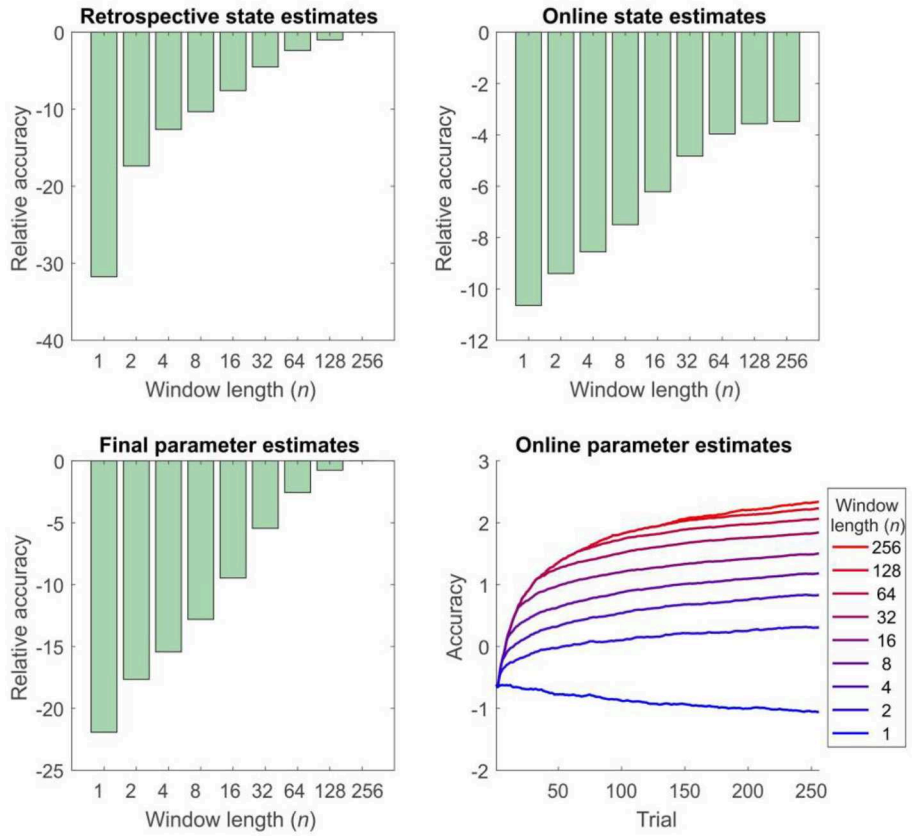


FIGURE 5 | Accuracy of inference and learning on the reversal task for agents using different window lengths, averaged across 1,000 simulations (see "Simulations" for more details). Accuracy is quantified as the log likelihood assigned to the true sequence of states or the true parameter value by the agent, averaged across simulations. Top left panel: accuracy of retrospective state estimation relative to the performance of an offline agent. Accuracy increases with window length, becoming identical for online and offline agents with the same effective window length (256 trials). Top right panel: accuracy of online (filtered) state estimates relative to the filtered state estimates of an offline agent. Accuracy increases with window length, but never becomes equivalent to that of an offline agent. This difference reflects the fact that the parameter estimates of the online agent only use observations made up to the present time, rather than on the entire data set (In other words, $p(\theta|o_{1:t}, \lambda)$ at trial t rather than $p(\theta|o_{1:T}, \lambda)$). This can be thought of as a cost of online inference. Bottom left panel: accuracy of final parameter estimates relative to the performance of an offline agent. Accuracy progressively increases with window length, becoming equivalent for online and offline agents with the same effective window length. Bottom right panel: average accuracy of parameter estimates across trials. Accuracy of parameter estimation increases with window length, and these differences progressively appear as the session goes on. (Absolute values of the accuracy measure are difficult to interpret here, but the relative accuracy of the difference agents is meaningful).

work, we plan to address this explicitly, both through simulations and experimental work.

The specific retrospective inference model we describe here differs importantly from previous approaches to modeling probabilistic reversal tasks (Hampton et al., 2006) and change point detection more generally (Wilson et al., 2010; Radillo et al., 2017) in two key ways, first through the fact that we allow for parameter learning (though see Radillo et al., 2017), and second, because we simulate agents that are able to update beliefs about past states. Both these processes are important for normative behavior, and it will be important to establish how closely human performance across a number of domains reflects this. Retrospective inference has also been considered in the context of reinforcement learning (Moran et al., 2019), and we will explore how to approach similar reward learning problems using our probabilistic framework in future. Similar ideas have also been explored in the context of active inference and planning (Friston et al., 2017; Kaplan and Friston, 2018), although these have not explored effects on learning.

Our approach differs importantly from models such as the hierarchical Gaussian filter (Mathys et al., 2011), which use higher-level variables operating at longer time scales to provide an implicit time window, but do not make postdictive inferences of the sort discussed here [In fact, retrospective inference has the potential to improve accuracy on tasks involving tracking of higher order variables like volatility (Behrens et al., 2007; Mathys et al., 2011), which is a promising area for future study]. A closer analogy can be drawn with generalized filtering approaches (Friston, 2008), which infer both on the current state and its derivatives (rate of change, acceleration and so on), and require a finite window of data to perform updates. This similarity is something we intend to return to in future work.

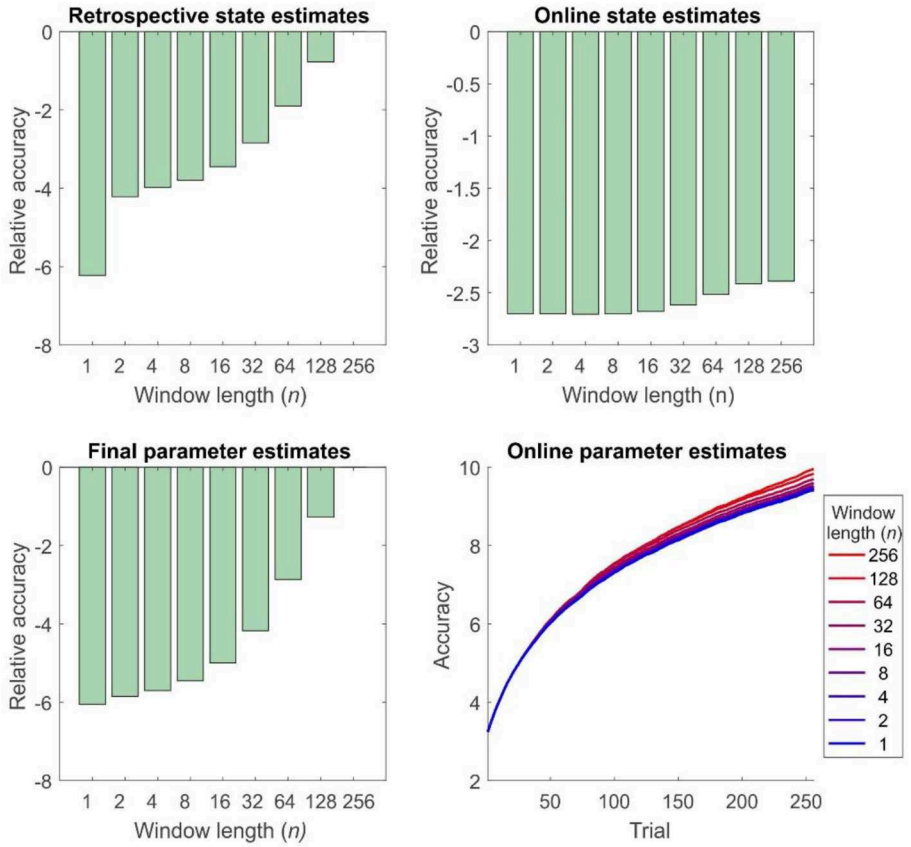


FIGURE 6 | Accuracy of inference and learning for random HMMs (see "Simulations" for more details). In general these mirror the results for the reversal task (**Figure 5**), but the quantitative differences are smaller, perhaps reflecting the greater number of states and parameters to be estimated (Accuracy is quantified as the log likelihood assigned to the true sequence of states or the true parameter value by the agent, averaged across simulations). Top left panel: accuracy of state estimation increases with window length, becoming identical for online and offline agents with the same effective window length (256 trials). Top right panel: accuracy of online (filtered) state estimates relative to the filtered state estimates of an offline agent. Accuracy increases with window length, but never becomes equivalent to that of an offline agent. Bottom left panel: accuracy of final parameter estimates relative to the performance of an offline agent. As the window length employed increases, so does accuracy, becoming equivalent for online, and offline agents with the same effective window length. Bottom right panel: average accuracy of parameter estimates across trials. Accuracy of parameter estimation increases with window length, and these differences progressively appear as the session goes on.

Retrospective inference provides a natural explanation for a number of "postdictive" phenomena in perception, in which perception of an event is influenced by other events that only occur afterwards (Eagleman and Sejnowski, 2000; Shimojo, 2014). A classic example of this is the color phi phenomenon (Kolers and von Grünau, 1976). Here, two differently colored dots are briefly displayed to the subject at different spatial locations. If the interval between the flashes is sufficiently short, subjects report perceiving a single moving dot, rather than two separate dots. Critically, they also perceive the color of the dot as changing during motion, meaning that they perceive the second color as occurring before it is presented on screen. This means that information about the color of the second dot has somehow been propagated backwards in (perceptual) time. That such postdictive phenomena might be explained by smoothing has previously been pointed out by Rao et al. (2001), but our proposal builds on this by suggesting a limited window of updating, as well as highlighting the importance of such belief updating for learning.

The existence of postdictive perceptual phenomena (among other considerations) have led to what is often called the "multiple drafts" account of consciousness (Dennett and Kinsbourne, 1992), in which the contents of conscious are subject to continual revision in the light of new information (at short timescales, at least), and what subjects report is critically dependent upon when they are asked. For example, in the color phi experiment, subjects' reported perceptual experience would differ if they were asked to report it before the second dot is shown, as opposed to when they are asked to report it afterwards. This accords extremely well with our proposal (at least if we make the further supposition that the contents of consciousness can, in some sense, be identified with the outcome of optimal perceptual inference). Under FRI (unlike filtering), reported perceptual experience will be critically dependent on when the report is made, since online retrospective inference makes beliefs

time-dependent. In other words, my belief about what happened at time *t* may be different depending on whether you ask me for it at time *t+1* or time *t+10*. This means that FRI has the potential to provide the computational underpinning of a "probabilistic multiple drafts" model of perceptual experience.

One intriguing possibility raised by FRI is that different individuals might perform retrospective belief updating to different extents, either on particular tasks or in general, and that this might partially explain between-subject differences in performance on particular tasks (see FitzGerald et al., 2017 for evidence of this). Such differences might even help explain facets of psychopathology (Montague et al., 2012). For example, impaired learning due to reduced or absent retrospection might lead to the tendency to form delusional beliefs (Hemsley and Garety, 1986; Corlett et al., 2004; Adams et al., 2013). For example, say someone looked at you in an unusual way—making you feel they were spying on you—but then subsequently ignored you: if you could not use the latter information to revise your initial suspicion, you would be more likely to become paranoid about that person. This idea is supported by the finding of altered neuronal responses in subjects with delusions (as compared with healthy controls) during performance of a retrospective belief updating task (Corlett et al., 2007), and is something we intend to return to in future.

Implementing retrospective inference also has important implications for neurobiology. In particular, since agents need to be able to dynamically update beliefs about past states, they are required to store explicit, ordered representations of the past, and it should be possible to find evidence of this in appropriate neuronal structures (Pezzulo et al., 2014) (For some evidence of this, see Corlett et al., 2004). Intriguingly, this fits extremely well with an extensive literature on hippocampal function (Fortin et al., 2002; Jensen and Lisman, 2005; Pastalkova et al., 2008; Lehn et al., 2009; Penny et al., 2013), a finding supported by the results of our previous study, which found a relation between depth of retrospective processing and gray matter density in the hippocampus (FitzGerald et al., 2017). On the further supposition that retrospective inference is implemented using filtering and smoothing as described above, this leads to the hypothesis that forward and backward sweeps through recently encountered states, as are known to occur in the hippocampus (Diba and Buzsáki, 2007; Pastalkova et al., 2008; Davidson et al., 2009; Wikenheiser and Redish, 2013) may play a key role in retrospective belief updating. What is less clear, at present, is how to implement retrospective inference within established, neurobiologically-grounded accounts of probabilistic inference in the brain (Friston, 2005; Ma et al., 2006; Aitchison and Lengyel, 2016)—though see (Friston et al., 2017) for related suggestions. This is an extremely important question, and one we intend to return to in future work.

Probabilistic models of cognition are an enormously exciting tool for understanding the complex workings of the mind and brain (Clark, 2012; Friston et al., 2013; Pouget et al., 2013; Aitchison and Lengyel, 2016). The ideas we propose represent a development of such approaches to encompass inference about states in the past, as well as the present. On the further hypothesis that the depth of processing employed is flexible and tailored to the demands of a particular problem or environment, such retrospective processing can also be linked to broader notions of bounded rationality (Simon, 1972; Gigerenzer and Goldstein, 1996; Ortega et al., 2015). We show, through simulations of simple environments, that even a limited degree of retrospection can yield significantly more accurate beliefs about both time-varying states and time-invariant parameters, and thus has the potential to support more adaptive, successful behavior to justify its extra resource costs. This makes it a plausible strategy for real, biological agents to employ FRI makes both behavioral and neuronal predictions in a number of contexts and thus naturally suggests further avenues for exploration in future work.

DATA AVAILABILITY STATEMENT

Matlab code for the modeling described is available at https://github.com/thbfitz/retro-inf.

AUTHOR CONTRIBUTIONS

TF conceived the study, created the models, and performed the simulations. WP contributed ideas and code. HB and RA contributed ideas. All authors contributed to writing the manuscript.

FUNDING

TF was supported by a European Research Council (ERC) Starting Grant under the Horizon 2020 program (Grant Agreement 804701). This manuscript has been released as a Pre-Print at https://www.biorxiv.org/content/10.1101/569574v2.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00002/full#supplementary-material

REFERENCES

Acuña, D. E., and Schrater, P. (2010). Structure learning in human sequential decision-making. *PLoS Comput. Biol.* 6:e1001003. doi: 10.1371/journal.pcbi.1001003

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., and Friston, K. J. (2013). The computational anatomy of psychosis. *Front. Psychiatry* 4:47. doi: 10.3389/fpsyt.2013.00047

Aitchison, L., and Lengyel, M. (2016). The hamiltonian brain: efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS Comput. Biol.* 12:e1005186. doi: 10.1371/journal.pcbi.1005186

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Mach. Learn.* 50, 5–43. doi: 10.1023/A:1020281327116

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* 1:0064. doi: 10.1038/s41562-017-0064

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of

markov chains. *Ann. Math. Stat.* 41, 164–171. doi: 10.1214/aoms/1177697196

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference* (Ph.D. thesis). University College London, London, United Kingdom.

Behrens, T. E., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221. doi: 10.1038/nn1954

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112, 859–877. doi: 10.1080/01621459.2017.1285773

Braun, D. A., Mehring, C., and Wolpert, D. D. M. (2010). Structure learning in action. *Behav. Brain Res.* 206, 157–165. doi: 10.1016/j.bbr.2009.08.031

Chen, B., and Tugnait, J. K. (2001). Tracking of multiple maneuvering targets in clutter using IMM/JPDA filtering and fixed-lag smoothing. *Automatica* 37, 239–249. doi: 10.1016/S0005-1098(00)00158-8

Clark, A. (2012). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Cohn, S. E., Sivakumaran, N. S., Todling, R., Cohn, S. E., Sivakumaran, N. S., and Todling, R. (1994). A fixed-lag kalman smoother for retrospective data assimilation. *Mon. Weather Rev.* 122, 2838–2867. doi: 10.1175/1520-0493(1994)122<;2838:AFLKSF>;2.0.CO;2

Corlett, P. R., Aitken, M. R., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A., et al. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron* 44, 877–888. doi: 10.1016/S0896-6273(04)00756-1

Corlett, P. R., Murray, G. K., Honey, G. D., Aitken, M. R., Shanks, D. R., Robbins, W., et al. (2007). Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain* 130, 2387–2400. doi: 10.1093/brain/awm173

Costa, V. D., Tran, V. L., Turchi, J., and Averbeck, B. B. (2015). Reversal learning and dopamine: a bayesian perspective. *J. Neurosci.* 35, 2407–2416. doi: 10.1523/JNEUROSCI.1989-14.2015

Davidson, T. J., Kloosterman, F., and Wilson, M. A. (2009). Hippocampal replay of extended experience. *Neuron* 63, 497–507. doi: 10.1016/j.neuron.2009.07.027

Dennett, D. C., and Kinsbourne, M. (1992). Time and the observer: the where and when of consciousness in the brain. *Behav. Brain Sci.* 15, 183–201. doi: 10.1017/S0140525X00068229

Diba, K., and Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nat. Neurosci.* 10, 1241–1242. doi: 10.1038/nn1961

Eagleman, D. M., and Sejnowski, T. J. (2000). Motion integration and postdiction in visual awareness. *Science* 287, 2036–2038. doi: 10.1126/science.287.5460.2036

FitzGerald, T. H., Hämmerer, D., Friston, K. J., Li, S.-C., and Dolan, R. J. (2017). Sequential inference as a mode of cognition and its correlates in fronto-parietal and hippocampal brain regions. *PLoS Comput. Biol.* 13:e1005418. doi: 10.1371/journal.pcbi.1005418

Fortin, N. J., Agster, K. L., and Eichenbaum, H. B. (2002). Critical role of the hippocampus in memory for sequences of events. *Nat. Neurosci.* 5, 458–462. doi: 10.1038/nn834

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Friston, K. (2008). Variational filtering. *Neuroimage* 41, 747–766. doi: 10.1016/j.neuroimage.2008.03.017

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Front. Hum. Neurosci.* 7:598. doi: 10.3389/fnhum.2013.00598

Friston, K., Trujillo-Barreto, N., and Daunizeau, J. (2008). DEM: a variational treatment of dynamic systems. *Neuroimage* 41, 849–885. doi: 10.1016/j.neuroimage.2008.02.054

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Gigerenzer, G., and Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychol. Rev.* 103, 650–669. doi: 10.1037/0033-295X.103.4.650

Glaze, C. M., Kable, J. W., and Gold, J. I. (2015). Normative evidence accumulation in unpredictable environments. *Elife* 4:e08825. doi: 10.7554/eLife.08825.019

Gregory, R. L. (1980). Perceptions as hypotheses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 290, 181–197. doi: 10.1098/rstb.1980.0090

Hampton, A. N., Bossaerts, P., and O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* 26, 8360–8367. doi: 10.1523/JNEUROSCI.1010-06.2006

Hassabis, D., Kumaran, D., and Maguire, E. A. (2007). Using imagination to understand the neural basis of episodic memory. *J. Neurosci.* 27, 14365–14374. doi: 10.1523/JNEUROSCI.4549-07.2007

Hemsley, D. R., and Garety, P. A. (1986). The formation of maintenance of delusions: a bayesian analysis. *Br. J. Psychiatry* 149, 51–56. doi: 10.1192/bjp.149.1.51

Jensen, O., and Lisman, J. E. (2005). Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends Neurosci.* 28, 67–72. doi: 10.1016/j.tins.2004.12.001

Kaplan, R., and Friston, K. J. (2018). Planning and navigation as active inference. *Biol. Cybern.* 112, 323–343. doi: 10.1007/s00422-018-0753-2

Kolers, P. A., and von Grünau, M. (1976). Shape and color in apparent motion. *Vision Res.* 16, 329–335. doi: 10.1016/0042-6989(76)90192-9

Lehn, H., Steffenach, H.-A., van Strien, N. M., Veltman, D. J., Witter, M. P., and Håberg, A. K. (2009). A specific role of the human hippocampus in recall of temporal sequences. *J. Neurosci.* 29, 3475–3484. doi: 10.1523/JNEUROSCI.5370-08.2009

Lieder, F., and Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychol. Rev.* 124, 762–794. doi: 10.1037/rev0000075

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat. Neurosci.* 9, 1432–1438. doi: 10.1038/nn1790

Mackay, D. J. C. (1997). *Ensemble Learning for Hidden Markov Models*. Technical report, Cavendish Laboratory, University of Cambridge.

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* 5:39. doi: 10.3389/fnhum.2011.00039

Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends Cogn. Sci.* 16, 72–80. doi: 10.1016/j.tics.2011.11.018

Moore, J. B. (1973). Discrete-time fixed-lag smoothing algorithms. *Automatica* 9, 163–173. doi: 10.1016/0005-1098(73)90071-X

Moran, R., Keramati, M., Dayan, P., and Dolan, R. J. (2019). Retrospective model-based inference guides model-free credit assignment. *Nat. Commun.* 10:750. doi: 10.1038/s41467-019-08662-8

Ortega, P. A., Braun, D. A., Dyer, J., Kim, K.-E., and Tishby, N. (2015). Information-theoretic bounded rationality. *arXiv* 1512.06789.

Pastalkova, E., Itskov, V., Amarasingham, A., and Buzsaki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science* 321, 1322–1327. doi: 10.1126/science.1159775

Penny, W. D., Zeidman, P., and Burgess, N. (2013). Forward and backward inference in spatial cognition. *PLoS Comput. Biol.* 9:e1003383. doi: 10.1371/journal.pcbi.1003383

Pezzulo, G., van der Meer, M. A., Lansink, C. S., and Pennartz, C. M. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends Cogn. Sci.* 18, 647–657. doi: 10.1016/j.tics.2014.06.011

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* 16, 1170–1178. doi: 10.1038/nn.3495

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286. doi: 10.1109/5.18626

Radillo, A. E., Veliz-Cuba, A., Josić, K., and Kilpatrick, Z. P. (2017). Evidence accumulation and change rate inference in dynamic environments. *Neural Comput.* 29, 1561–1610. doi: 10.1162/NECO_a_00957

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rao, R. P. N., Eagleman, D. M., and Sejnowski, T. J. (2001). Optimal smoothing in visual motion perception. *Neural Comput.* 13, 1243–1253. doi: 10.1162/08997660152002843

Russell, S., and Wefald, E. (1991). Principles of metareasoning. *Artif. Intell.* 49, 361–395. doi: 10.1016/0004-3702(91)90015-C

Sarkka, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge: Cambridge University Press.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., and Szpunar, K. K. (2012). The future of memory: remembering, imagining, and the brain. *Neuron* 76, 677–694. doi: 10.1016/j.neuron.2012.11.001

Schlagenhauf, F., Huys, Q. J., Deserno, L., Rapp, M. A., Beck, A., Heinze, H. J., et al. (2014). Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *Neuroimage* 89, 171–180. doi: 10.1016/j.neuroimage.2013.11.034

Shimojo, S. (2014). Postdiction: its implications on visual awareness, hindsight, and sense of agency. *Front. Psychol.* 5:196. doi: 10.3389/fpsyg.2014.00196

Simon, H. A. (1972). "Theories of bounded rationality," in *Decision and Organization*, eds C. McGuire and R. Radner (Amsterdam: North-Holland), 161–176.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

Tervo, D. G. R., Tenenbaum, J. B., and Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Curr. Opin. Neurobiol.* 37, 99–105. doi: 10.1016/j.conb.2016.01.014

von Helmholtz, H. (1867). *Handbuch der Physiologischen Optik*. Leipzig: Leopold Voss.

Wan, E. A., Van Der Merwe, R., and Nelson, A. T. (1999). "Dual estimation and the unscented transformation," in *Proceedings of the 12th International Conference on Neural Information Processing Systems* (Cambridge, MA: MIT Press), 666–672.

Wikenheiser, A. M., and Redish, A. D. (2013). The balance of forward and backward hippocampal sequences shifts across behavioral states. *Hippocampus* 23, 22–29. doi: 10.1002/hipo.22049

Wilson, R. C., Nassar, M. R., and Gold, J. I. (2010). Bayesian online learning of the hazard rate in change-point problems. *Neural Comput.* 22, 2452–2476. doi: 10.1162/NECO_a_00007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 28 February 2020
doi: 10.3389/frai.2020.00005

Information Theoretic Characterization of Uncertainty Distinguishes Surprise From Accuracy Signals in the Brain

Leyla Loued-Khenissi[1]* *and Kerstin Preuschoff*[2,3]

[1] Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [2] Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland, [3] Geneva Finance Research Institute, University of Geneva, Geneva, Switzerland

Uncertainty presents a problem for both human and machine decision-making. While utility maximization has traditionally been viewed as the motive force behind choice behavior, it has been theorized that uncertainty minimization may supersede reward motivation. Beyond reward, decisions are guided by belief, i.e., confidence-weighted expectations. Evidence challenging a belief evokes surprise, which signals a deviation from expectation (stimulus-bound surprise) but also provides an information gain. To support the theory that uncertainty minimization is an essential drive for the brain, we probe the neural trace of uncertainty-related decision variables, namely confidence, surprise, and information gain, in a discrete decision with a deterministic outcome. Confidence and surprise were elicited with a gambling task administered in a functional magnetic resonance imaging experiment, where agents start with a uniform probability distribution, transition to a non-uniform probabilistic state, and end in a fully certain state. After controlling for reward expectation, we find confidence, taken as the negative entropy of a trial, correlates with a response in the hippocampus and temporal lobe. Stimulus-bound surprise, taken as Shannon information, correlates with responses in the insula and striatum. In addition, we also find a neural response to a measure of information gain captured by a confidence error, a quantity we dub accuracy. BOLD responses to accuracy were found in the cerebellum and precuneus, after controlling for reward prediction errors and stimulus-bound surprise at the same time point. Our results suggest that, even absent an overt need for learning, the human brain expends energy on information gain and uncertainty minimization.

Keywords: uncertainty, information theory, surprise, confidence, probabilistic brain, fMRI, decision-making

OPEN ACCESS

Edited by:
Thomas Parr,
University College London,
United Kingdom

Reviewed by:
Giovanni Pezzulo,
Italian National Research Council, Italy
Jakub Limanowski,
University College London,
United Kingdom

***Correspondence:**
Leyla Loued-Khenissi
lkhenissi@gmail.com

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

Received: *30 October 2019*
Accepted: *03 February 2020*
Published: *28 February 2020*

Citation:
Loued-Khenissi L and Preuschoff K (2020) Information Theoretic Characterization of Uncertainty Distinguishes Surprise From Accuracy Signals in the Brain. Front. Artif. Intell. 3:5. doi: 10.3389/frai.2020.00005

1. INTRODUCTION

Uncertainty is a feature of an agent's interaction with the environment that is both pervasive and unavoidable. Its ubiquity therefore demands a place in an agent's decision-making calculus. But uncertainty emerges in different forms during a decision, each of which can be uniquely susceptible to dysfunction. During an initial deliberation phase, for instance, agents form a belief on a decision's outcome, which is graded by confidence (Kepecs and Mainen, 2012). An outcome that

challenges beliefs yields surprise (Hsia, 1991; Nour et al., 2018; Munnich and Ranney, 2019). Both confidence and surprise relate to uncertainty in the environment but their characterization remains a topic of debate (Itti and Baldi, 2009; Baldi and Itti, 2010; Munnich et al., 2019). Surprise may generate at least two quantities: one relating to an event's frequency (stimulus-bound surprise), and another back-propagating information gain that fine-tunes initial beliefs (model update) (Lorini and Castelfranchi, 2007; Itti and Baldi, 2009; Faraji et al., 2018). These two quantities together make up the uncertainty defined in the Free Energy Principle (Friston, 2010), whose minimization is theorized to be the brain's primary purpose (Schwartenbeck et al., 2015) and comprises a compelling theoretical framework for brain function. Questions on the neural characterization of different forms of uncertainty persist for both confidence (Pouget et al., 2016) and surprise (Munnich and Ranney, 2019). Current studies investigating uncertainty in the brain often rely on the notion of a Bayesian brain (Friston, 2012), where a probabilistic model of the world is built (the prior) and subsequently updated (posterior) through repeated interactions with the environment. In this paper, we seek to disentangle different aspects of uncertainty, namely confidence, as well as the dual facets of surprise, by applying a parsimonious, information theoretic model to BOLD response signals in a functional magnetic resonance imaging experiment. A neural response to these quantities would lend support for their emergence in the decision-making process.

1.1. Confidence

Human confidence is often thought of as a feeling but its mathematical definition has been extensively used in the fields of statistics and economics (Dominitz and Manski, 2004; Cesarini et al., 2006) and has more recently attracted interest in the neuroscience of decision-making (Kepecs et al., 2008; Kiani and Shadlen, 2009; Rolls et al., 2010; De Martino et al., 2013). Most studies on confidence in decision-making employ a subjective measure of post-decision confidence, obtained via self-report or inferred from reaction time (Kepecs and Mainen, 2012). Confidence arising prior to a decision outcome by contrast is a form of prediction uncertainty (Meyniel et al., 2015), or the second-order uncertainty coupled to a first-order expectation (Preuschoff et al., 2008a,b) and can be represented by the inverse variance (precision) (Yeung and Summerfield, 2014; Pouget et al., 2016) or the negative entropy of a probability distribution. Confidence is thought to weight both belief and the impact of its eventual violation: the more precise the prediction, the more significant its associated error (Feldman and Friston, 2010; Kwisthout et al., 2017). Neuroimaging studies on prediction uncertainty, specifically entropy and variance, have uncovered related BOLD responses in the hippocampus (Strange et al., 2005; Harrison et al., 2006; Davis et al., 2012), the striatum and insula (Preuschoff et al., 2006, 2008b; Mohr et al., 2010). Although confidence figures prominently in predictive processing theory (Friston et al., 2012; Barrett and Simmons, 2015), comparatively few neuroimaging studies have probed its unique contribution and neural representation. As confidence can confer an affective state (Sanders et al., 2016), it may correlate to anterior insular responses, and as it depends on prior knowledge, it may also relate to memory regions, such as the hippocampus and temporal lobe. Here, we seek a neural response to confidence as formalized by an information theoretic quantity, namely the negative entropy of a probability distribution, when an agent formulates an expectation.

1.2. Surprise

The error related to prediction uncertainty is commonly cast as surprise (Hayden et al., 2011; Preuschoff et al., 2011). The problem of surprise in both artificial intelligence and cognitive neuroscience hinges on its definition, which in turn opens a fraught discourse on its putative purpose (Munnich et al., 2019). From a phenomenological perspective, surprise is an organism's response to an unexpected change in her environment. Formal accounts of the phenomenon include Shannon surprise (Shannon, 1948); Bayesian surprise (Itti and Baldi, 2009); a predictive coding account of surprise [as absolute prediction error (Pearce and Hall, 1980) or risk prediction error (Preuschoff et al., 2011)]. These accounts share common features but are not perfectly correlated and, in some instances, can yield diverging values (Baldi and Itti, 2010). Broadly speaking, all but Bayesian Surprise can be considered "stimulus-bound" surprise, although both risk and absolute prediction error further integrate the value of an event, while Shannon Surprise is invariant to the latter. Itti and Baldi (2009) posit that an event can only be surprising if there is *post-hoc* evidence of learning; that is, the relevance of an event elicits surprise, not merely its improbability (Weaver; Faraji et al., 2018). Itti and Baldi formally distinguish Shannon surprise as stimulus-bound surprise and Bayesian surprise, an information gain represented by a Kullback-Leibler divergence (DKL) between prior and posterior beliefs (Itti and Baldi, 2009). They further argue that it is Bayesian Surprise that constitutes true surprise. However, one can argue that a rare event, formalized by Shannon surprise, is always relevant. The Free Energy framework (Friston, 2009) accounts for these distinct formulations of surprise by allowing for both stimulus-bound surprise and model update to constitute a measure of uncertainty (Free Energy), whose minimization is theorized to drive an agent (Schwartenbeck et al., 2015). In the brain, surprise as expectation violation correlates with BOLD responses in the salience network, including the anterior cingulate cortex and anterior insula (Uddin, 2014; Gogolla, 2017). Here, we seek to replicate previous results found in relation to stimulus-bound surprise specifically by applying an information theoretic account to the BOLD response, as the latter does not integrate the value of an event as risk and absolute prediction error do.

1.3. Information Gain

An unexpected outcome presents an opportunity to learn but more fundamentally, a chance to acquire knowledge. An intelligent agent should therefore exploit unexpected events so as to gain information. Information gain is commonly taken to be the Kullback–Leibler divergence, or relative entropy, which conforms to the notion of a Bayesian brain (Knill and Pouget, 2004) and therefore, implicitly, an assumption that certitude is never encountered (Basieva et al., 2017).

However, an argument can be made that, in some instances and at higher levels of brain hierarchy, humans rely on approximate solutions and therefore can experience certitude. When a model cannot be further updated, or, in Markovian terms, when an agent reaches a terminal state, information gained from an event can be characterized as the difference between the truth (outcome) and the degree of prior belief (confidence), or absolute entropy (Shannon, 1948). What bridges the gap between belief and knowledge is an information gain and can be cast as an accuracy term. While accuracy is commonly taken as the difference between observed and (average) expected outcomes, we take it to be the difference between observed and the upper limit of expected outcomes (confidence). Thus, information gain may arise even if the model space is confined to one decision and can be defined for cases in which predictions are perfect, or outcomes are certain, as the self-evidence of a prediction (Parr et al., 2018), or the confirmation of a belief. For instance, suppose an agent invests in a given company's stock, estimating both it's future stock price and a confidence interval on that estimate. The agent wakes several years later to find the stock price has shot up suddenly, exceeding her expectations. The difference between the estimated and true stock price prompts a reward prediction error; the rarity of the event prompts surprise; and the discrepancy between the agent's confidence and the true outcome, or how far off the mark the agent was, represents a form of accuracy, or information gain. As in confidence, Bayesian formalization of information gain has gained considerable traction in recent years, but it can be argued that purely information theoretic accounts can simplify uncertainty quantification (Thornton, 2017). It is possible that the brain expends no resources on information gain if there is no future model to update however, a case can also be made for the curious brain, an information-hungry organism that collects and hoards evidence for possible future use. Here, we explore the neural response to a non-Bayesian information gain, which notably can be used in one-shot decisions.

1.4. Empirical Evidence of Stimulus-bound Surprise and Model Update

The dual aspect of surprise as both an alarm signal and a quantity of information is theoretically compelling, but less convincing in a human context. Stimulus-bound surprise necessarily calls on an autonomic response (Preuschoff et al., 2011), while an information gain need not. Several empirical studies have sought neural evidence of surprise's dual role. An examination of surprise models in P300 ERP signals finds Shannon information best explained data rather than a KL divergence, or a model that discounted forgetting across study blocks (Mars et al., 2008). Stimulus-bound rather than Bayesian surprise provided a better fit to the P300 ERP, widely viewed as a neural "surprise" signal, however, evidence of distinct neural systems correlating to stimulus-bound surprise and Bayesian surprise were found using fMRI (O'Reilly et al., 2013; Schwartenbeck et al., 2015; Kobayashi and Hsu, 2017). These studies suggest that, in humans (1) stimulus-bound surprise comprises a relevant phenomenon and that (2) a surprise-related learning signal also implicates a neural response. What remains unknown is whether a neural response reflecting information gain, distinct from a signed prediction error and stimulus-bound surprise, can be identified in the case of a one-shot decision process with a deterministic outcome where the Kullback–Leibler divergences cannot be computed. Such a signal can serve as a stand-in for subjective measures of post-decision confidence, bypassing report-related error and would also lend credence to the principle of uncertainty minimization as a primary neural drive.

In the following study, we examine three main questions in the context of value-based decision-making under uncertainty. We seek the neural representation of distinct but related uncertainty variables, notably confidence, surprise and accuracy. Specifically we hypothesize that (1) stimulus-bound surprise will elicit a BOLD response in the insula, striatum, anterior cingulate as in previous studies pertaining to error detection; (2) that confidence signals will be reflected in the insula, striatum and hippocampus, as entropy and risk have in other studies; (3) that accuracy signals will incur a unique BOLD response after accounting for reward prediction error and stimulus-bound surprise at the same time point. We test our hypotheses using fMRI within the context of a gambling paradigm that elicits both uncertainty predictions as well as their concomitant errors while controlling for reward, motivational, learning and motor effects. Capturing these quantities in the brain can inform on the human decision-making process, and notably provide guidance in where the process can fail. Several clinical populations show signs of dysfunctional decision-making (Pellicano and Burr, 2012; Limongi et al., 2018), yet the precise nature of these lapses in judgment remains difficult to quantify. By the same token, a more detailed description of the human decision-making process can guide efforts in artificial intelligence by providing more variables with which a machine can learn.

2. MATERIALS AND METHODS

To examine our question of interest, we re-analyzed data from an auditory gambling task performed during fMRI acquisition. In the previous study, we sought commonalities of uncertainty processing in perception and value-based decision making task (Loued-Khenissi et al., 2020).

2.1. Participants

Twenty-nine healthy participants (10 F, average age 25.13 years) completed the experiment. Participants were recruited via paper and online advertisements targeting the student populations of Ecole Polytechnique Fédérale de Lausanne and Université de Lausanne. Exclusion criteria included metal implants, previous psychiatric illness, and psychotropic drug use within the past year. Inclusion criteria included proficiency in English.

2.2. Behavioral Task

To induce our target uncertainty variables, we employed an auditory version of a gambling task that has previously yielded responses to both prediction uncertainty and surprise (Preuschoff et al., 2006, 2008b). In the task, participants were asked to bet on the outcome of a card game. Starting with an initial endowment of 25 CHF (25 USD), participants bet 1

CHF that a second card drawn from a deck of 10 cards would be higher or lower than a first card. Bets were placed prior to any card being sounded. After the bet, the two cards were revealed sequentially, with a time lag of 5.5 s between their sounding. After the first card, participants could compute their chance of winning (predicted reward), as well as a confidence in their trial outcome prediction (predicted uncertainty). Once the second card was revealed, participants could assess their errors in reward and uncertainty prediction. Following the second card's sounding, participants were asked to report the bet's outcome, as a means of controlling for attention. Onsets for Cards 1 and 2 were separated by 5.5 s intervals, to better differentiate hemodynamic response function peaks relating to predictive and outcome phases of decision-making. A random jitter of 2–5 s was included following each trial. Each round of the card game lasted 25 s. To control for fatigue and attention, a penalty of 25 c was included for each missed bet and each missed or incorrect report. Participants viewed a black fixation cross on a gray screen during the imaging session, while stimuli were presented in pre-recorded wav files transmitted to MR compatible headphones, using Mac OS's text to speech function (**Figure 1**). The experimental task was written in Matlab (Matlab and Statistics Toolbox Release 2013a, TheMathWorks, Inc., Natick, Massachusetts, United States) using the Psychophysics toolbox (Kleiner, 2010). Participants were paid for their time at the end of the experimental session; task-related payout was reserved for a subsequent second experimental session, to lower rates of attrition.

2.3. Imaging Procedure

All neuroimaging data were acquired on a Siemens 3T Prisma at the Centre Hospitalier Universitaire Vaudois. Parameters for the EPI sequence were: 2D EPI, Multi-Echo sequence (3 echo times), 3 x 3 x 2.5 mm resolution, FOV = 192 mm; FA = 90 degrees, slice TR = 80 ms; TE = (17.4; 35.2; 53 ms); base resolution 64 mm; 34 slices; volume TR = 2.72 s; parallel acceleration mode = GRAPPA, with an acceleration factor = 2. At the end of the experimental session, anatomical T1 images were acquired with the following parameters: T1 MPRAGE, 1x1x1 mm resolution; FOV = 256 mm; slice TR/TE = 2 ms/2.39 ms; FA = 9 degrees; base resolution = 256 mm).

2.4. Image Preprocessing

Functional scans were preprocessed and analyzed using SPM12. Echo volumes were first summed to obtain one scan per TR. We then performed slice-timing correction and generated voxel displacement maps (VDM) to apply to functional volumes. Volumes were warped and realigned to the mean functional image using a 6 parameter (translations and rotations in space), rigid-body transformation to correct motion artifacts, before

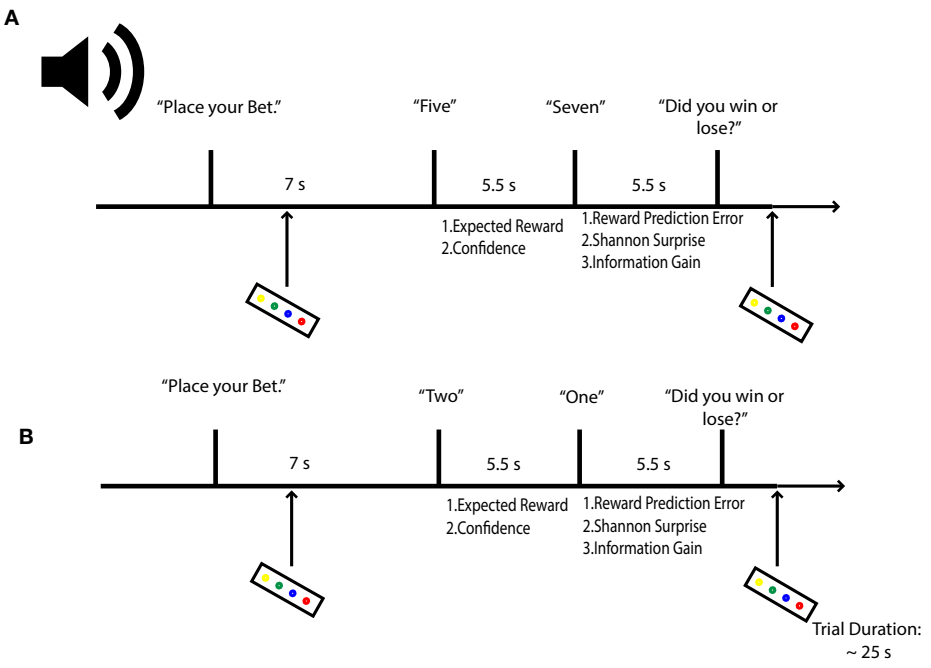


FIGURE 1 | Probabilistic Gambling Task. Participants were asked to place bets on whether a second card draw from a deck of 10 would be higher or lower in value than a first card. Bets were placed before either card was revealed. Participants estimate their reward (expected value) and confidence (expected uncertainty) in the bet outcome after hearing card 1. After hearing card 2, agents can compute their reward prediction error; their stimulus-bound surprise and also their confidence error or information gain. Let us assume in the above example that a participant bets the second card will be lower. In **(A)**, confidence in the outcome will be low, given that the first card is a five; surprise is also expected to be low when card 2 is revealed, but information will be high, as the second card can take on several states for each outcome (1–4 for lower values, 6–10 for higher values) relative to the first card. In **(B)**, a participant should be confident that she will lose, as there is only one card out of a possible 9 that can deliver a win; therefore, when the second card yields the improbably one, surprise is expected to be high. Concomitantly, information gain is expected to be low, as confidence in the outcome had to be high.

being bias-field corrected. Then individual T1 volumes were co-registered to the mean functional image using a rigid body model, estimated with mutual information. The T1 image was then segmented (6 class tissue probability maps) and normalized to MNI space using unified segmentation (Ashburner and Friston, 2005). These normalization parameters were then applied to functional volumes. Volumes were then smoothed with a Gaussian kernel of 8 mm FWHM.

2.5. Mathematical Models

The task employed was designed to evoke probabilistic inferences in participants. The decision variables derived below are based on the probability distribution of winning (or losing) a gamble. Our computational model for reward prediction at card 1 reflects the average expected reward given the bet placed (higher or lower), and card 1's value (Preuschoff et al., 2011). The reward prediction error at card 2 reflects the trial outcome (win or loss) minus the reward prediction. Confidence is taken as the negative entropy H of outcome probability distributions after Card 1. This quantity is always negative and tends, when H = 0, toward 0. While negative entropy and inverse variance are often used interchangeably to quantify uncertainty and are numerically equivalent for most cases in our dataset, the inverse variance is necessarily undefined when $\rho = 0$. One could approximate such "infinite" confidence by setting $\rho(0) = \epsilon$, however resultant values will 1) depend on ϵ; 2) yield a value for infinite confidence that is not ordinal to other values of confidence (**Figure 2**). At card 2, Shannon information quantifies stimulus-bound surprise, as the negative log of the probability of the observed outcome, x, given the bet placed, b and the value of Card 1, c. Finally, information gain was captured by the difference between the maximal value of confidence (certitude), minus confidence at Card 1. We take this maximal confidence to be 0; the information gain is thus always $=< 0$, as it is the DKL; to differentiate this quantity from other forms of information gain, we call it accuracy. Because our task begins with an equal probability of outcome and ends with a terminal state that is independent of prior and future trial outcomes, we do not expect any learning to occur between trials. The trial begins with a flat prior and ends with a pseudo-deterministic outcome. Therefore, trials are assumed to be independent.

$$H = -p_{win} \cdot \log_2(p_{win}) - (1 - p_{win}) \cdot \log_2(1 - p_{win}) \quad (1)$$

$$Confidence = -H \quad (2)$$

$$Surprise = -\log_2 \cdot (p(outcome|bet, card1)) \quad (3)$$

$$Accuracy = 0 - Confidence \quad (4)$$

2.6. Imaging Analysis

We performed a model-based analysis on our functional neuroimaging data. Specifically, we parametrically modulated onsets of interest by mathematical quantities described below. At the subject level, we constructed a general linear model including one regressor for sound activation (following onset of instructions to place the bet and to report the gamble outcome, modeled by a Dirac function); one regressor for motor response (including onsets for bet placement and outcome report, modeled as a Dirac function); a regressor for onsets of the first card's presentation (modeled as 5.5.s boxcar function), parametrically modulated first by reward prediction, followed by confidence; and a regressor for onsets of card 2's presentation (modeled as 5.5.s boxcar function), parametrically modulated first by the reward prediction error; second, by stimulus-bound surprise; and finally by an accuracy term. Parametric modulators were serially orthogonalized in the

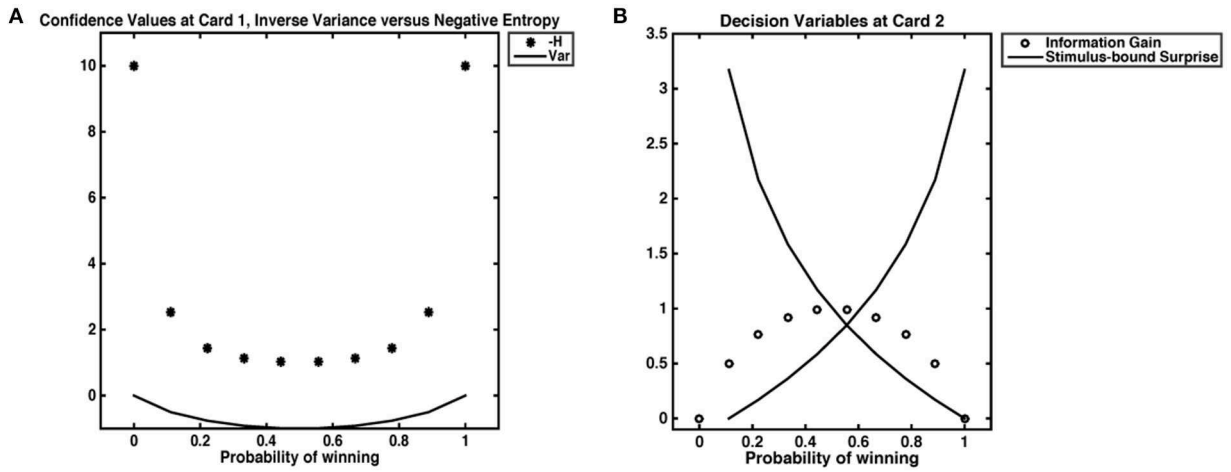


FIGURE 2 | Decision variables. **(A)** Confidence as negative entropy or inverse variance. When outcomes are certain, neither inverse variance nor negative entropy are defined. However, approximating negative entropy by 0 yields a value that is ordinal to the next highest levels of confidence, while approximating 0 variance with an ϵ of 0.001 gives a value that does not scale with others confidence values. **(B)** We show the relationship between stimulus-bound surprise and information gain. Confidence is at its lowest when the probability of a win is 0.5; in such an instance, an agent has the most information to gain but does not experience the least (or most) surprise. Highest surprise is reserved for instances where confidence was high, such as cases where the probability of a win is 0.9; in such an instance, a loss would necessarily incur high Shannon surprise.

order described above, ensuring that related BOLD responses to specific decision-making variables reflect that variable's unique contribution to the signal. Also included in the model were 6 motion-related regressors of no interest. We note that BOLD responses to expected reward and reward prediction errors were not of primary interest to our study; they are nonetheless included in the general linear model so as to account for their unique contribution to the BOLD response, thereby allowing for the isolation of uncertainty-related variables. Onsets were convolved with the canonical hemodynamic response function. The time-series was high-pass filtered (128 s); autocorrelation was modeled by an AR(1) function. We performed *t*-tests at the single subject level on confidence, Shannon surprise and accuracy regressors. Individual contrast images were then pooled as estimates in a random-effects model. At the group level, we conducted non-parametric tests using the SnPM13 toolbox (10 000 permutations, variance smoothing = 8 mm).

3. RESULTS

3.1. Behavioral Results

Twenty-five participants were included in the analysis. Behavioral data was not acquired for the first three participants. A fourth participant showed an error rate in excess of 30% (tallied from missed bets and reports, as well as incorrect reports) and was excluded from further analysis. Average task-related payout was 29.57 CHF; across all sessions and subjects, payoffs were in the range of 13-39 CHF. As the task designed included a truly random presentation of card pairs, we performed *post-hoc* analyses on potential differences for several variables of interest across sessions. We performed an *F*-test to determine if any one session contained more of one type of card value for card 1 and found no significant differences across sessions ($F = 0$, $p = 0.996$). We then performed tests on the mean differences of higher bets and lower bets across sessions and found no significant differences ($F = 0.19$, $p = 0.8324$ and $F = 0.2$, $p = 0.8204$, respectively), suggesting participants did not "switch" strategies across sessions. We also analyzed bet choices within blocks, by summing bet switches following a loss with bet persistence after a win, to assess the possible influence of prior bet outcomes. We find participants chose "non-strategic" bets more often ($t = -3.01$, $p = 0.0035$, df = 74), suggesting participants did not attempt to "learn" from previous outcomes. We also found a significant difference in bet choices with a higher likelihood for selecting a higher bet in all sessions ($F = 34.69, p < 0.001$).

3.2. Neuroimaging Results

We report results of voxels that remain significant when corrected for multiple comparisons, at a threshold of $p = 0.05$, FWE corrected at the whole brain level. Voxels were localized with the use of the Neuromorphometrics toolbox (Neuromorphometrics, Inc).

3.2.1. Confidence at Card 1

We performed a *t*-test on the onset of card 1's sounding for the prediction phase of the trial (duration = 5.5 s), parametrically modulated by confidence. Confidence here is orthogonal to reward prediction (experienced during the same time interval). We find a significant cluster in the right hippocampus; bilateral middle frontal gyrus; left supramarginal gyrus; right angular gyrus; right middle temporal gyrus; left superior temporal gyrus; and left inferior frontal gyrus. (**Figure 3**; **Table 1**).

3.2.2. Stimulus-Bound Surprise at Card 2

A *t*-test was performed on the onset of Card 2, parametrically modulated by stimulus-bound surprise of the trial for the duration between card 2's sounding and the outcome report (5.5 s). Significant clusters were found in expected regions, notably in the dorsal striatum (left putamen, right caudate); bilateral inferior frontal gyrii, extending into the anterior insula; left posterior cingulate cortex; bilateral medial temporal gyrii; and left supramarginal gyrus (**Figure 4**; **Table 2**).

3.2.3. Accuracy at Card 2

A *t*-test was performed on the onset of Card 2, parametrically modulated by the accuracy of a trial, for a duration of 5.5 s. This quantity was included in the GLM as a third parametric modulator to Card 2's onset, following reward prediction error and stimulus-bound surprise. Significant voxels were found in the left supramarginal gyrus; bilateral; precuneus; bilateral cerebellum (exterior); and left central operculum (**Figure 5**; **Table 3**).

3.2.4. Learning Across Trials

The experimental paradigm employed assumes no learning occurs across trials. Where there may be a learning effect is in the unlikely event that a subject counts card pairs as they are presented, because each possible card pair is only presented once. Should a subject deduce that each card pair is only presented once and also retain card pair values in memory as the experiment proceeds, we may expect the model space to expand to the experimental session. We nonetheless controlled for the possibility that a subject counted cards during the experimental sessions by designing a second GLM that differed from that described above only in swapping information with a Bayesian update measure. We computed this Bayesian update measure by employing a Dirichlet counting process, as per Strange et al. (2005), where wins were counted across a session, and included this measure of learning or divergence in a general linear model as a parametric regressor at Card 2. No significant voxels emerged, even when lowering the threshold to $p = 0.05$, uncorrected.

$$p_{win_i} = \frac{\sum_1^i Wins + 1}{\sum_1^i Outcomes + 1} \tag{5}$$

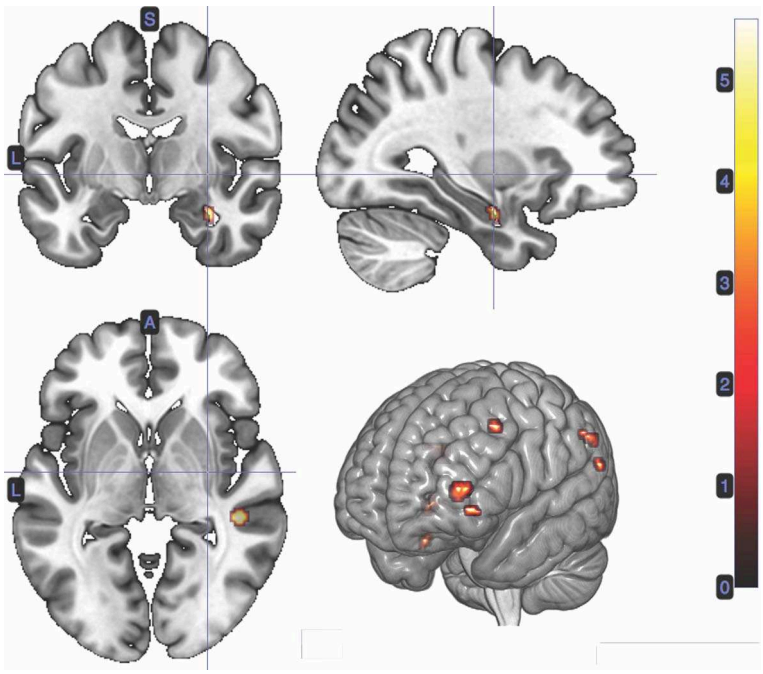


FIGURE 3 | Statistical non-parametric map of significant clusters correlating to confidence in the interval between Card 1 and Card 2. Maps were thresholded with $p = 0.05$, FWE-corrected for multiple comparisons. The colorbar indicates t-values.

4. DISCUSSION

The results above show that (1) confidence, as negative entropy, correlates with the hippocampus, a region previously linked to uncertainty processing; (2) stimulus-bound surprise elicits activity in the insula and striatum, replicating previous studies; (3) accuracy, as a measure of information gain sampled at the same timepoint as stimulus-bound surprise, elicits a BOLD response in distinct regions, namely the cerebellum and precuneus. By using a formal account of all three measures while controlling for reward-related decision variables as well as task-related phenomena, such as overt learning and motor action, we link confidence, surprise and information gain to distinct neural correlates using information theoretic accounts. The emergence of a BOLD response for these three quantities underlines uncertainty's importance in human decision-making and lends empirical support to the principles of both uncertainty minimization and evidence maximization in brain function (Hohwy, 2012; Fiorillo, 2017; Pezzulo and Friston, 2019). Moreover, the localization of neural responses to surprise and information gain closely mirror a recent fMRI study investigating the similar questions but with the use of a Bayesian model (Kobayashi and Hsu, 2017).

TABLE 1 | Statistics and locations of significant ($p = 0.05$, FWE-corrected) peaks and clusters related to confidence at Card 1.

		Confidence				
k	**FWE**	***T***	***x***	***y***	***z***	**Region**
65	0.0022	5.61	−42	48	6	L MFG
32	0.0028	5.56	46	−32	−2	R MTG
32	0.006	5.33	−58	−50	38	L Supramarginal Gyrus
96	0.006	5.32	58	−54	28	R Angular Gyrus
23	0.0162	4.95	−44	18	44	L MFG
11	0.0188	4.9	32	−8	−22	R Hippocampus
12	0.0208	4.87	−48	44	−4	L IFG
9	0.0354	4.64	−60	−56	22	L Superior Temporal Gyrus
2	0.0368	4.63	−58	−42	40	L Supramarginal Gyrus
1	0.0464	4.54	62	−48	18	R Angular Gyrus

4.1. Confidence

In our study, both the hippocampus and temporal gyrus correlate with confidence measures, in line with our hypothesis. Our results support the notion that confidence occupies a particular role in decision-making variables (Friston, 2018; Kiani and Shadlen, 2009; Insabato et al., 2010; Pouget et al., 2016). Confidence measures in human studies often suffer from being a self-reported, subjective measure assessed *post-hoc*. Here, we examine an objective form of confidence, captured by the negative entropy computed during a passive, predictive phase of an event's outcome. As prediction is theorized to arise from integrating an incoming stimulus into prior knowledge (Clark, 2013), memory regions should be implicated in this phase of decision-making. Previous studies have found a BOLD response in the hippocampus for related measures of prediction uncertainty such as variability (Rigoli et al., 2019) and entropy

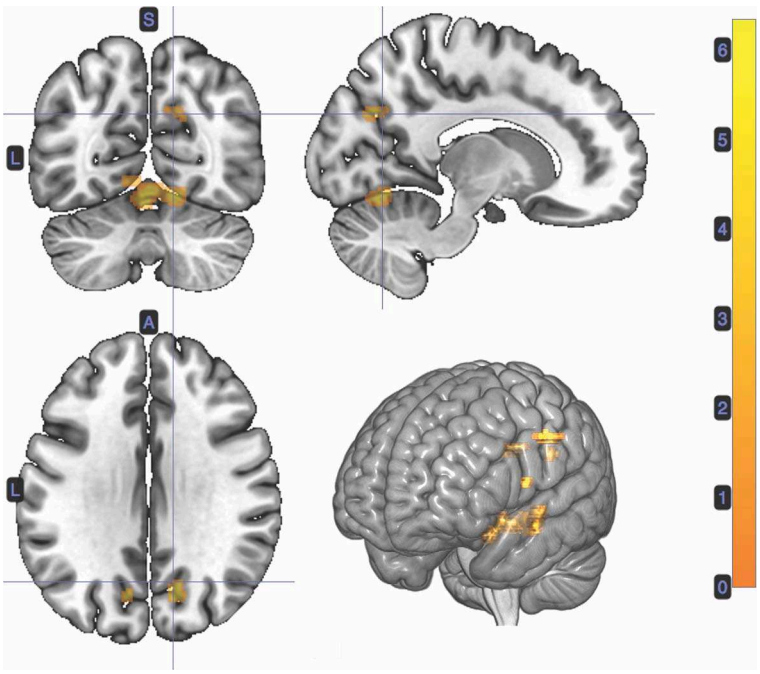


FIGURE 4 | Statistical non-parametric map of significant clusters correlating to stimulus-surprise at trial outcome. Stimulus-bound surprise represented a second parametric modulator of Card 2's event onset, after accounting for the reward prediction error. Maps were thresholded with $p = 0.05$, FWE-corrected for multiple comparisons. The colorbar indicates *t*-values.

TABLE 2 | Statistics and locations of significant ($p = 0.05$, FWE-corrected) peaks and clusters related to stimulus-bound surprise at Card 2.

Stimulus-bound Surprise						
k	**FWE**	***T***	***x***	***y***	***z***	**Region**
655	0.0008	6.24	−22	−2	8	L Putamen
-	0.0012	5.85	−22	8	−6	–
-	0.002	5.57	−34	18	2	L Ains
738	0.0014	5.69	18	10	12	R Caudate
-	0.0022	5.51	44	20	−12	–
-	0.0022	5.49	24	−4	6	R Putamen
272	0.002	5.59	58	18	12	R IFG/Ains
-	0.0118	5.04	52	14	18	–
-	0.0126	5.02	52	30	16	–
69	0.0022	5.47	−62	−52	4	L MTG
38	0.007	5.19	0	−30	28	L PCG
108	0.0096	5.14	58	−56	6	R MTG
-	0.014	4.97	54	−46	12	–
110	0.0106	5.1	−50	38	6	L IFG
-	0.0156	4.91	−44	34	12	–
59	0.0128	5.01	-58	−52	24	L SupraMarginalGyrus
55	0.015	4.94	56	−36	−2	R MTG
57	0.016	4.9	−56	16	12	L IFG
-	0.021	4.82	−50	10	14	–

Clusters with more than one significant peak in the same region are indicated with a dash.

(Strange et al., 2005; Harrison et al., 2006) but here we explicitly find hippocampal responses for confidence, and not entropy or risk. Further, by using negative entropy rather than inverse variance, we divorce this quantity from the expected mean; that is, confidence is invariant to the value of the prediction. Our results further add to the current body of knowledge pertaining to brain correlates of confidence because we employ a whole-brain rather than ROI-based analysis. Other areas correlating with confidence include parietal regions, namely bilateral angular and supramarginal gyri. Angular gyri have previously been implicated in decision-making under uncertainty in humans (Symmonds et al., 2011; Studer et al., 2014). In monkeys, parietal neurons have previously been found to encode perceptual confidence using an evidence accumulation model (drift diffusion) in rhesus monkeys (Kiani and Shadlen, 2009). Finally, parietal lesions in humans have been found to leave recollection unaltered, but to specifically impair memory confidence (Simons et al., 2010). It is noteworthy that none of the studies above explicitly model confidence as negative entropy, but nonetheless yield similar neuroanatomical correlates. While the parietal lobe was not a primary focus of our hypothesis on the neural correlates of confidence, results from the extant literature validate our use of an information theoretic model of confidence.

4.2. Stimulus-Bound Surprise

We find evidence of stimulus-bound surprise in the (posterior) cingulate cortex and anterior insula, regions thought to signal

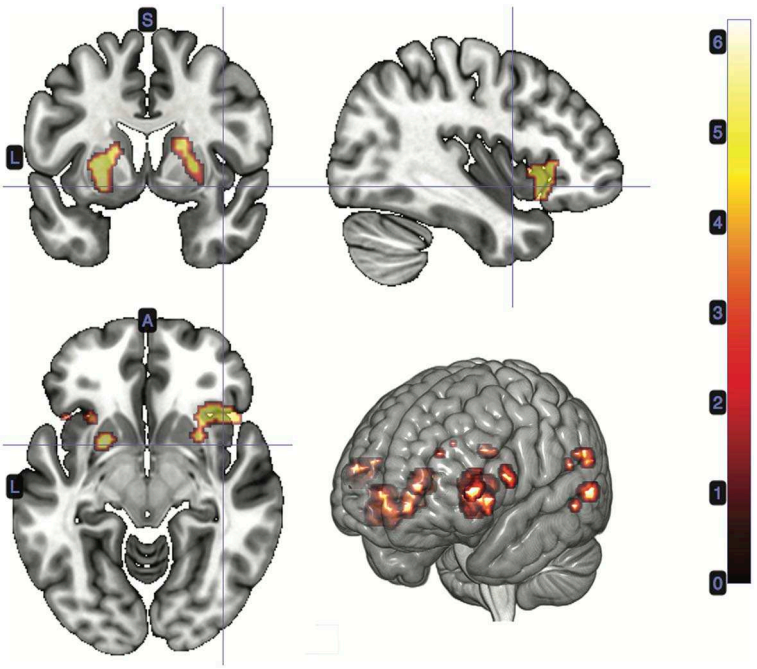


FIGURE 5 | Statistical non-parametric map of significant clusters correlating to information gain at trial outcome. Information gain represented a third parametric modulator of Card 2's event onset, after accounting for the reward prediction error and stimulus-bound surprise. Maps were thresholded with $p = 0.05$, FWE-corrected for multiple comparisons. The colorbar indicates *t*-values.

error detection and conflict (Ullsperger et al., 2010); and the striatum, all regions previously implicated in studies on surprise (Preuschoff et al., 2011; Kobayashi and Hsu, 2017) but not found in other studies investigating both stimulus-bound surprise and information gain (O'Reilly et al., 2013; Schwartenbeck et al., 2015). Our results reaffirm the neural relevance of event improbability decoupled from the nature of the event (gain or loss) and by extension, the likely behavioral pertinence of such outcomes. Here, by controlling for the contributions of both the reward prediction error and information gain to the BOLD response at the outcome of a trial, we can confidently assert that our measure of surprise captures error-detection free of a hedonic component. Significant responses in the temporal lobe, a memory region, further add credence to the predictive processing framework. Stimulus-bound surprise can only occur when an event is compared to a prior expectation, a state of affairs that necessitates a memory component.

TABLE 3 | Statistics and locations of significant ($p = 0.05$, FWE-corrected) peaks and clusters related to Information Gain (Accuracy) at Card 2.

Information Gain (Accuracy)						
k	**FWE**	***T***	***x***	***y***	***z***	**Region**
68		0.0002 6.34	−38	−38	38	L Supramarginal Gyrus
44	0.09	5.1	12	−70	32	R Precuneus
240	0.01	5.04	−10	-60	−10	L Cerebellum
-	0.0136	4.93	16	−64	−12	–
-	0.0158	4.87	24	−58	−20	–
17	0.0288	4.68	−12	−72	28	L Precuneus
6	0.0386	4.56	−42	−14	12	L Central Operculum/Posterior Insula
1	0.0486	4.46	−44	−28	40	L Post CentralGyrus

Clusters with more than one significant peak in the same region are labeled with a dash.

4.3. Model Update, Learning, and Accuracy

Evidence of learning can best reflect an information gain. However, no learning is expected to occur in our task, and this by design. All trials start with an equal probability of winning, so no strategizing can occur and outcomes do not depend on previous trials. We nonetheless captured signals related to a quantity of information gain by measuring maximal minus predicted confidence, or absolute entropy (Shannon, 1948). To distinguish this quantity from a model update (O'Reilly et al., 2013) we call this error term *accuracy*. Absent such a signal, we can hypothesize that no information has been gained, which suggests an agent was certain in the predictive phase of a decision. Accuracy was reflected in the cuneus and cerebellum. The cuneus has previously been implicated in learning rates (Payzan-LeNestour et al., 2013) and belief updating (Kobayashi and Hsu, 2017), in line with results in our study and has also been implicated in perceptual evidence accumulation (Ploran et al., 2011; FitzGerald et al., 2015), however this region also correlated with stimulus-bound surprise in another fMRI study (O'Reilly et al., 2013). The cerebellum on the other hand showed the strongest response to information gain. While a role for the cerebellum has been hypothesized in learning (Doya, 2000; Friston and Buzsáki, 2016) and inferential processes (Blackwood

et al., 2004; Friston and Buzsáki, 2016), it is not commonly viewed as a decision-making hub. Of note is the lack of BOLD response in the cingulate cortex, which contrasts with results found by O'Reilly et al. in their study (2013). The absence of a BOLD response in the cingulate cortex, a region commonly linked to conflict (Botvinick, 2007) underlines the quality of information gain, in that it need not stem from incongruence but more fundamentally as an acquisition of knowledge, even while being a "prediction error." Our results underline the inherent value information has (Friston et al., 2012), for the brain would not expend energy on a response otherwise. The brain may collect seemingly useless information, for a potential future. The implication of information collection is not trivial: it supports the notion that an agent may want to maximize her entropy (Schwartenbeck et al., 2013) and in so doing "seek" surprise (Clark, 2018), or a state of expecting the unexpected (Sun et al., 2011. Indeed, those individuals with stronger signals relating to information gain may be cast as more adventurous, or risk-seeking (Kruschwitz et al., 2012).

4.4. Hypothesized Disruptions of the Probabilistic Brain

Elucidation of uncertainty decision-variables can help identify specific components of dysfunctional decision-making and learning, particularly in patient populations (Parr et al., 2018). Isolating a neural response to confidence alone, for instance, may help shed light on aberrant decision-making. A compromised ability to compute confidence may lie at the heart of pathologies such as obsessive-compulsive disorder (OCD) (Hermans et al., 2008; Vaghi et al., 2017) and anxiety (Grupe and Nitschke, 2013; Carleton, 2016). Therefore, one could probe a patient's response to confidence in the hippocampus to determine if it deviates from a healthy range. Both repetitive actions and negative outlooks (expecting the worst) may increase confidence, and therefore minimize (unpleasant) surprise in OCD and anxiety patients (Hein et al., 2019), respectively; but increasing confidence would also erroneously minimize information gain (Kwisthout et al., 2017) and therefore accuracy. While these strategies are maladaptive, they are not irrational; framing them in the context of aberrant computations offers a way to identify the specific sub-process causing distress (Parr et al., 2018). Probabilistic computation may also be compromised in autism (Sinha et al., 2014; Van de Cruys et al., 2014); and schizophrenia (Silverstein et al., 2017). For instance, autistic individuals overestimate the volatility of an uncertain environment (Lawson et al., 2017). A disorder where stimulus-bound surprise is not computed may result in apathy and flattened affect, a common symptom in schizophrenic patients. On the other hand, an inflated stimulus-bound surprise could overwhelm an agent, which may be a feature of autism. Difficulty acquiring information specifically by discounting the accuracy term above could impede an agent's change in belief. Similarly, too large an information gain signal could indicate false belief formation (Schwartenbeck et al., 2015). Therefore, the neural processing of each of the quantities probed above may contribute to a specific dysfunction in behavior. Simulations of agents with specific deficits can be conducted to predict pathological symptoms of different psychiatric disorders.

4.5. Uncertainty in Man and Machines

The findings above also impact questions in artificial intelligence (Macedo and Cardoso, 2001; Lorini and Castelfranchi, 2006; Lorini and Piunti, 2007). If artificial intelligence is modeled after human behavior (Lake et al., 2017) then formalizing and finding evidence of the processes deployed in human intelligence offers a more precise template to reproduce. The utility in endowing a an intelligent agent with uncertainty and model update computation is clear. Less convincing is the need to encode all forms of uncertainty-related variables. Humans need stimulus-bound surprise, as it prompts a fight or flight response, presumably in the face of death: updating a model may well be irrelevant in such a case, or at the very least, secondary. In machines however, a model update may be necessary and sufficient, while stimulus-bound surprise may be surperfluous. Another consideration with respect to artificial modeling of surprise is the inclusion of its affective component. Hedonic components of surprise, such as positive and negative valence, can be accounted for in the sign of the reward prediction error. However, human surprise is also tinged with a range of other graded emotions: joy, disappointment, disgust, horror, anger, awe and fear (Braem et al., 2015). One could engage in a thought experiment to identify cases when an artificial agent may need to "feel" different hues of surprise-specific emotion. There may be no concrete purpose in endowing an artificial agent with the capacity to encode awe, for instance.

4.6. Conclusions

Our aim was to employ information theory to model and decompose uncertainty signals in the brain. Studies investigating the probabilistic brain have primarily exploited Bayesian models (Knill and Pouget, 2004; Friston, 2012) however as seen in the study above, such models may not easily accommodate certitude or one-shot decisions. While our work cannot identify causal relationships between external stimuli and recorded BOLD signals, we nonetheless find a relationship between the two. Significant brain responses that correlate to specific formal accounts suggest such calculations are being performed. In finding distinct responses to confidence, surprise and information gain, we highlight the importance of uncertainty integration to the brain. In identifying a neural correlate of information gain for a discrete decision in particular we: 1) offer an alternative to the Bayesian Surprise model of the latter; 2) show that the brain seeks to maximize evidence even when there is no obvious reason to do so. The implications of our results may help refine efforts to model intelligent agents and provide specific measures to identify and quantify decision-making deficits in clinical populations.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available (1) At the time of data collection, participants were not asked permission for dissemination of their data.

(2) There is no known way to completely anonymize neuroimaging data, as images allow for a crude form of facial reconstruction.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Swissethics (EC Vaud). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LL-K designed the study, acquired the data, performed the analysis, and wrote the paper. KP designed the study and wrote the paper.

FUNDING

This work was supported by the Swiss National Science Foundation (320030L_135687).

REFERENCES

Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *NeuroImage* 26, 839–851. doi: 10.1016/j.neuroimage.2005.02.018

Baldi, P., and Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Netw.* 23, 649–666. doi: 10.1016/j.neunet.2009.12.007

Barrett, L. F., and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* 16:419. doi: 10.1038/nrn3950

Basieva, I., Pothos, E., Trueblood, J., Khrennikov, A., and Busemeyer, J. (2017). Quantum probability updating from zero priors (by-passing Cromwells rule). *J. Math. Psychol.* 77, 58–69. doi: 10.1016/j.jmp.2016.08.005

Blackwood, N., ffytche, D., Simmons, A., Bentall, R., Murray, R., and Howard, R. (2004). The cerebellum and decision making under uncertainty. *Cognit. Brain Res.* 20, 46–53. doi: 10.1016/j.cogbrainres.2003.12.009

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognit. Affect. Behav. Neurosci.* 7, 356–366. doi: 10.3758/CABN.7.4.356

Braem, S., Coenen, E., Bombeke, K., van Bochove, M. E., and Notebaert, W. (2015). Open your eyes for prediction errors. *Cogn. Affect. Behav. Neurosci.* 15, 374–380. doi: 10.3758/s13415-014-0333-4

Carleton, R. N. (2016). Into the unknown: a review and synthesis of contemporary models involving uncertainty. *J. Anxiety Disord.* 39, 30–43. doi: 10.1016/j.janxdis.2016.02.007

Cesarini, D., Sandewall, r., and Johannesson, M. (2006). Confidence interval estimation tasks and the economics of overconfidence. *J. Econ. Behav. Organ.* 61, 453–470. doi: 10.1016/j.jebo.2004.10.010

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenom. Cogn. Sci.* 17, 521–534. doi: 10.1007/s11097-017-9525-z

Davis, T., Love, B. C., and Preston, A. R. (2012). Striatal and hippocampal entropy and recognition signals in category learning: simultaneous processes revealed by model-based fMRI. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 821–839. doi: 10.1037/a0027865

De Martino, B., Fleming, S. M., Garrett, N., and Dolan, R. J. (2013). Confidence in value-based choice. *Nat. Neurosci.* 16, 105–110. doi: 10.1038/nn.3279

Dominitz, J., and Manski, C. F. (2004). How should we measure consumer confidence? *J. Econ. Perspect.* 18, 51–66. doi: 10.1257/0895330041371303

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Curr. Opin. Neurobiol.* 10, 732–739. doi: 10.1016/S0959-4388(00)00153-7

Faraji, M., Preuschoff, K., and Gerstner, W. (2018). Balancing new against old information: the role of puzzlement surprise in learning. *Neural Comput.* 30, 34–83. doi: 10.1162/neco_a_01025

Feldman, H., and Friston, K. (2010). Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* 4:215. doi: 10.3389/fnhum.2010.00215

Fiorillo, C. D. (2017). Neuroscience: rationality, uncertainty, dopamine. *Nat. Hum. Behav.* 1, 1–2. doi: 10.1038/s41562-017-0158

FitzGerald, T. H. B., Moran, R. J., Friston, K. J., and Dolan, R. J. (2015). Precision and neuronal dynamics in the human posterior parietal cortex during evidence accumulation. *NeuroImage* 107, 219–228. doi: 10.1016/j.neuroimage.2014.12.015

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cognit. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage* 62, 1230–1233. doi: 10.1016/j.neuroimage.2011.10.004

Friston, K. (2018). Does predictive coding have a future? *Nat Neurosci.* 21, 1019–1021. doi: 10.1038/s41593-018-0200-7

Friston, K., Adams, R., and Montague, R. (2012). What is value accumulated reward or evidence? *Front. Neurorobot.* 6:11. doi: 10.3389/fnbot.2012.00011

Friston, K., and Buzsáki, G. (2016). The functional anatomy of time: what and when in the brain. *Trends Cognit. Sci.* 20, 500–511. doi: 10.1016/j.tics.2016.05.001

Gogolla, N. (2017). The insular cortex. *Curr. Biol.* 27, R580–R586. doi: 10.1016/j.cub.2017.05.010

Grupe, D. W., and Nitschke, J. B. (2013). Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat. Rev. Neurosci.* 14, 488–501. doi: 10.1038/nrn3524

Harrison, L. M., Duggins, A., and K.j, F. (2006). Encoding uncertainty in the hippocampus. *Neural Networks.* 19:535. doi: 10.1016/j.neunet.2005.11.002

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., and Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *J. Neurosci.* 31, 4178–4187. doi: 10.1523/JNEUROSCI.4652-10.2011

Hein, T. P., Weber, L. A., Fockert, J. d., and Ruiz, M. H. (2019). State anxiety biases estimates of uncertainty during reward learning in volatile environments. *bioRxiv* 809749. doi: 10.1101/809749

Hermans, D., Engelen, U., Grouwels, L., Joos, E., Lemmens, J., and Pieters, G. (2008). Cognitive confidence in obsessive-compulsive disorder: Distrusting perception, attention and memory. *Behav. Res. Therap.* 46, 98–113. doi: 10.1016/j.brat.2007.11.001

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Front. Psychol.* 3:96. doi: 10.3389/fpsyg.2012.00096

Hsia, Y.-T. (1991). "Belief and surprise-a belief-function formulation," in *Uncertainty Proceedings 1991.* 165–173. doi: 10.1016/B978-1-55860-203-8.50025-5

Insabato, A., Pannunzi, M., Rolls, E. T., and Deco, G. (2010). Confidence-Related Decision Making. *J. Neurophysiol.* 104, 539–547. doi: 10.1152/jn.01068.2009

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vis. Res.* 49, 1295–1306. doi: 10.1016/j.visres.2008.09.007

Kepecs, A., Uchida, N., Zariwala, H. A., and Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455, 227–231. doi: 10.1038/nature07200

Kepecs, A., and Mainen Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 1322–1337. doi: 10.1098/rstb.2012.0037

Kiani, R., and Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* 324, 759–764. doi: 10.1126/science.1169405

Kleiner, M. (2010). "Visual stimulus timing precision in Psychtoolbox-3: Tests, pitfalls and solutions," in *Perception 39 ECVP Abstract Supplement* (Lausanne), 189.

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Kobayashi, K., and Hsu, M. (2017). Neural mechanisms of updating under reducible and irreducible uncertainty. *J. Neurosci.* 37, 6972–6982. doi: 10.1523/JNEUROSCI.0535-17.2017

Kruschwitz, J. D., Simmons, A. N., Flagan, T., and Paulus, M. P. (2012). Nothing to lose: processing blindness to potential losses drives thrill and adventure seekers. *NeuroImage* 59, 2850–2859. doi: 10.1016/j.neuroimage.2011.09.048

Kwisthout, J., Bekkering, H., and van Rooij, I. (2017). To be precise, the details dont matter: on predictive processing, precision, and level of detail of predictions. *Brain Cognit.* 112, 84–91. doi: 10.1016/j.bandc.2016.02.008

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci.* 40:e253. doi: 10.1017/S0140525X16001837

Lawson, R. P., Mathys, C., and Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nat. Neurosci.* 20, 1293–1299. doi: 10.1038/nn.4615

Limongi, R., Bohaterewicz, B., Nowicka, M., Plewka, A., and Friston, K. J. (2018). Knowing when to stop: aberrant precision and evidence accumulation in schizophrenia. *Schizophrenia Res.* 197, 386–391. doi: 10.1016/j.schres.2017.12.018

Lorini, E., and Castelfranchi, C. (2006). The unexpected aspects of surprise. *Int. J. Patt. Recogn. Artif. Intell.* 20, 817–833. doi: 10.1142/S0218001406004983

Lorini, E., and Castelfranchi, C. (2007). The cognitive structure of surprise: looking for basic principles. *Topoi* 26, 133–149. doi: 10.1007/s11245-006-9000-x

Lorini, E., and Piunti, M. (2007). "The benefits of surprise in dynamic environments: from theory to practice," in *Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science*, eds A. C. R. Paiva, R. Prada, and R. W. Picard (Berlin; Heidelberg: Springer), 362–373.

Loued-Khenissi, L., Pfeuffer, A., EinhÃuser, W., and Preuschoff, K. (2020). Anterior insula reflects surprise in value-based decision-making and perception. *NeuroImage* 210:116549. doi: 10.1016/j.neuroimage.2020.116549

Macedo, L., and Cardoso, A. (2001). "Modeling forms of surprise in an artificial agent," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 23.

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., and Bestmann, S. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *J. Neurosci.* 28, 12539–12545. doi: 10.1523/JNEUROSCI.2925-08.2008

Meyniel, F., Sigman, M., and Mainen, Z. F. (2015). Perspective confidence as bayesian probability : from neural origins to behavior. *Neuron* 88, 78–92. doi: 10.1016/j.neuron.2015.09.039

Mohr, P. N. C., Biele, G., and Heekeren, H. R. (2010). Neural processing of risk. *J. Neurosci.* 30, 6613–9. doi: 10.1523/JNEUROSCI.0003-10.2010

Munnich, E., and Ranney, M. A. (2019). Learning from surprise: harnessing a metacognitive surprise signal to build and adapt belief networks. *Topics Cognit. Sci.* 11, 164–177. doi: 10.1111/tops.12397

Munnich, E. L., Foster, M. I., and Keane, M. T. (2019). Editors introduction and review: an appraisal of surprise: tracing the threads that stitch it together. *Top. Cogn. Sci.* 11, 37–49. doi: 10.1111/tops.12402

Nour, M. M., Dahoun, T., Schwartenbeck, P., Adams, R. A., FitzGerald, T. H. B., Coello, C., et al. (2018). Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. *Proc. Natl. Acad. Sci. U.S.A.* 115, E10167–E10176. doi: 10.1073/pnas.1809298115

O'Reilly, J. X., Schuffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., and Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proc. Natl. Acad. Sci. U.S.A.* 110, E3660–E3669. doi: 10.1073/pnas.1305373110

Parr, T., Rees, G., and Friston, K. J. (2018). Computational Neuropsychology and Bayesian Inference. *Front. Hum. Neurosci.* 12:61. doi: 10.3389/fnhum.2018.00061

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., and O'Doherty, J. P. (2013). The neural representation of unexpected uncertainty during value-based decision making. *Neuron* 79, 191–201. doi: 10.1016/j.neuron.2013.04.037

Pearce, J. M., and Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* 87, 532–552. doi: 10.1037/0033-295X.87.6.532

Pellicano, E., and Burr, D. (2012). When the world becomes too real: a Bayesian explanation of autistic perception. *Trends Cognit. Sci.* 16, 504–510. doi: 10.1016/j.tics.2012.08.009

Pezzulo, G., and Friston, K. J. (2019). The value of uncertainty: an active inference perspective. *Behav. Brain Sci.* 42:e47. doi: 10.1017/S0140525X18002066

Ploran, E. J., Tremel, J. J., Nelson, S. M., and Wheeler, M. E. (2011). High quality but limited quantity perceptual evidence produces neural accumulation in frontal and parietal cortex. *Cereb Cortex* 21, 2650–2662. doi: 10.1093/cercor/bhr055

Pouget, A., Drugowitsch, J., and Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* 19:366. doi: 10.1038/nn.4240

Preuschoff, K., Bossaerts, P., and Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51, 381–390. doi: 10.1016/j.neuron.2006.06.024

Preuschoff, K., Quartz, S., and Bossaerts, P. (2008a). Markowitz in the brain? *Rev. Econ. Politique* 118, 75–95. doi: 10.3917/redp.181.0075

Preuschoff, K., Quartz, S. R., and Bossaerts, P. (2008b). Human insula activation reflects risk prediction errors as well as risk. *J. Neurosci.* 28, 2745–2752. doi: 10.1523/JNEUROSCI.4286-07.2008

Preuschoff, K., 't Hart, B. M., and EinhÃuser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. *Front. Neurosci.* 5:115. doi: 10.3389/fnins.2011.00115

Rigoli, F., Michely, J., Friston, K. J., and Dolan, R. J. (2019). The role of the hippocampus in weighting expectations during inference under uncertainty. *Cortex* 115, 1–14. doi: 10.1016/j.cortex.2019.01.005

Rolls, E. T., Grabenhorst, F., and Deco, G. (2010). Decision-making, errors, and confidence in the brain. *J. Neurophysiol.* 104, 2359–2374.

Sanders, J. I., Hangya, B., and Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron* 90, 499–506. doi: 10.1016/j.neuron.2016.03.025

Schwartenbeck, P., FitzGerald, T., Dolan, R., and Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Front. Psychol.* 4:710. doi: 10.3389/fpsyg.2013.00710

Schwartenbeck, P., Fitzgerald, T. H., Mathys, C., Dolan, R., Kronbichler, M., and Friston, K. (2015). Evidence for surprise minimization over value maximization in choice behavior. *Sci. Rep.* 5, 1–14. doi: 10.1038/srep16575

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Silverstein, S. M., Wibral, M., and Phillips, W. A. (2017). Implications of Information Theory for Computational Modeling of Schizophrenia. *Comput. Psychiatry* 1, 82–101. doi: 10.1162/CPSY_a_00004

Simons, J. S., Peers, P. V., Mazuz, Y. S., Berryhill, M. E., and Olson, I. R. (2010). Dissociation between memory accuracy and memory confidence following bilateral parietal lesions. *Cereb Cortex* 20, 479–485. doi: 10.1093/cercor/bhp116

Sinha, P., Kjelgaard, M. M., Gandhi, T. K., Tsourides, K., Cardinaux, A. L., Pantazis, D., et al. (2014). Autism as a disorder of prediction. *Proc. Natl. Acad. Sci. U.S.A.* 111, 15220–15225. doi: 10.1073/pnas.1416797111

Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., and Friston, K. J. (2005). Information theory, novelty and hippocampal responses: Unpredicted or unpredictable? *Neural Netw.* 18, 225–230. doi: 10.1016/j.neunet.2004.12.004

Studer, B., Cen, D., and Walsh, V. (2014). The angular gyrus and visuospatial attention in decision-making under risk. *NeuroImage* 103, 75–80. doi: 10.1016/j.neuroimage.2014.09.003

Sun, Y., Gomez, F., and Schmidhuber, J. (2011). "Planning to be surprised: optimal Bayesian exploration in dynamic environments," in *Artificial General Intelligence, Lecture Notes in Computer Science*, eds J. Schmidhuber, K. R. ThÃrisson, and M. Looks (Berlin; Heidelberg: Springer), 41–51.

Symmonds, M., Wright, N. D., Bach, D. R., and Dolan, R. J. (2011). Deconstructing risk: separable encoding of variance and skewness in the brain. *NeuroImage* 58, 1139–1149. doi: 10.1016/j.neuroimage.2011.06.087

Thornton, C. (2017). Predictive processing simplified: The infotropic machine. *Brain Cogn.* 112, 13–24. doi: 10.1016/j.bandc.2016.03.004

Uddin, L. Q. (2014). Salience processing and insular cortical function and dysfunction. *Nat. Rev. Neurosci.* 16:55. doi: 10.1038/nrn3857

Ullsperger, M., Harsay, H. A., Wessel, J. R., and Ridderinkhof, K. R. (2010). Conscious perception of errors and its relation to the anterior insula. *Brain Struct. Funct.* 214, 629–643. doi: 10.1007/s00429-010-0261-1

Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., and De Martino, B. (2017). Compulsivity reveals a novel dissociation between action and confidence. *Neuron* 96, 348–354.e4. doi: 10.1016/j.neuron.2017.09.006

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de Wit, L., et al. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychol. Rev.* 121, 649–675. doi: 10.1037/a0037665

Weaver, W. Probability, rarity, interest, and surprise. *The Scientific Monthly* 67, 390–392.

Yeung, N., and Summerfield, C. (2014). "Shared mechanisms for confidence judgements and error detection in human decision making," in *The Cognitive Neuroscience of Metacognition*, eds S. M. Fleming and C. D. Frith (Berlin; Heidelberg: Springer), 147–167. doi: 10.1007/978-3-642-45190-4_7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

HYPOTHESIS AND THEORY
published: 09 June 2020
doi: 10.3389/frai.2020.00030

An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic Causation

*Adam Safron**

Indiana University, Bloomington, IN, United States

OPEN ACCESS

Edited by:
Maxwell James D. Ramstead,
McGill University, Canada

Reviewed by:
Adam B. Barrett,
University of Sussex, United Kingdom
Francis Fallon,
St. John's University, United States

***Correspondence:**
Adam Safron
asafron@gmail.com

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

Received: *16 December 2019*
Accepted: *03 April 2020*
Published: *09 June 2020*

Citation:
Safron A (2020) An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic Causation. Front. Artif. Intell. 3:30. doi: 10.3389/frai.2020.00030

The Free Energy Principle and Active Inference Framework (FEP-AI) begins with the understanding that persisting systems must regulate environmental exchanges and prevent entropic accumulation. In FEP-AI, minds and brains are predictive controllers for autonomous systems, where action-driven perception is realized as probabilistic inference. Integrated Information Theory (IIT) begins with considering the preconditions for a system to intrinsically exist, as well as axioms regarding the nature of consciousness. IIT has produced controversy because of its surprising entailments: quasi-panpsychism; subjectivity without referents or dynamics; and the possibility of fully-intelligent-yet-unconscious brain simulations. Here, I describe how these controversies might be resolved by integrating IIT with FEP-AI, where integrated information only entails consciousness for systems with perspectival reference frames capable of generating models with spatial, temporal, and causal coherence for self and world. Without that connection with external reality, systems could have arbitrarily high amounts of integrated information, but nonetheless would not entail subjective experience. I further describe how an integration of these frameworks may contribute to their evolution as unified systems theories and models of emergent causation. Then, inspired by both Global Neuronal Workspace Theory (GNWT) and the Harmonic Brain Modes framework, I describe how streams of consciousness may emerge as an evolving generation of sensorimotor predictions, with the precise composition of experiences depending on the integration abilities of synchronous complexes as self-organizing harmonic modes (SOHMs). These integrating dynamics may be particularly likely to occur via richly connected subnetworks affording body-centric sources of phenomenal binding and executive control. Along these connectivity backbones, SOHMs are proposed to implement turbo coding via loopy message-passing over predictive

(autoencoding) networks, thus generating maximum a posteriori estimates as coherent vectors governing neural evolution, with alpha frequencies generating basic awareness, and cross-frequency phase-coupling within theta frequencies for access consciousness and volitional control. These dynamic cores of integrated information also function as global workspaces, centered on posterior cortices, but capable of being entrained with frontal cortices and interoceptive hierarchies, thus affording agentic causation. Integrated World Modeling Theory (IWMT) represents a synthetic approach to understanding minds that reveals compatibility between leading theories of consciousness, thus enabling inferential synergy.

Keywords: consciousness, free energy principle, active inference, generative model, autonomy, integrated information theory, global workspace, autoencoder

INTRODUCTION AND BACKGROUND

Here, I introduce *Integrated World Modeling Theory (IWMT)* as a synthetic approach to understanding consciousness, using the *Free Energy Principle and Active Inference Framework (FEP-AI)* (Friston et al., 2006, 2017a; Friston, 2010) to combine multiple theories into a unified perspective. IWMT focuses on *Integrated Information Theory (IIT)* (Tononi, 2004; Tononi et al., 2016) and *Global Neuronal Workspace Theory (GNWT)* (Baars, 1993; Dehaene, 2014) as two of the most well-known theories of consciousness. Areas of agreement and disagreement between IIT and GNWT will be explored, as well as the extent to which points of contention might be productively addressed by situating these theories within FEP-AI. I then review the fundamentals of FEP-AI as a general systems theory, including points of intersection with IIT as an account of causal emergence. I then go on to discuss mechanistic and computational principles by which these theories can all be integrated using IWMT. In brief, IWMT states that consciousness may be what it is like to be processes capable of generating integrated models of systems and worlds with spatial, temporal, and causal coherence. IWMT further suggests that such coherence is only likely to be attainable for embodied agentic systems with controllers capable of supporting complexes of high degrees of integrated information, functioning as global workspaces and arenas for Bayesian model selection. Finally, I consider potential implications of these proposals with respect to the enduring problems of consciousness and artificial intelligence.

Toward Integration

How can physical systems generate subjective experiences? Can mental states function as causes, or are we mere automata? These perennial questions may finally be answerable with two unifying frameworks for understanding complex systems and minds: FEP-AI and IIT. These two meta-theoretical frameworks were developed in the context of understanding psychological and neurobiological phenomena, yet their implications are far more extensive. FEP-AI may be the first unified formalism and paradigm for the mind and life sciences, and IIT is one of the most widely known and technically detailed models of consciousness and informational synergy. FEP-AI describes what systems must be like in order to persist, and IIT describes what it means for systems to intrinsically exist as systems. Both FEP-AI and IIT constitute general systems theories with scopes transcending disciplinary boundaries, having relevance not only for the philosophy and science of mind but also for understanding all emergent complexity.

Here, I describe how these two frameworks complement each other as unified systems theories, and also show how FEP-AI allows IIT and GNWT to be combined into a synthetic framework for understanding consciousness: IWMT. This synthesis further attempts to characterize the nature of mental causation in terms of generalized Darwinism (Campbell, 2016) and thermodynamic work cycles, thus describing how conscious agency may be essential for understanding how flexible intelligence may be realized in biological (and potentially artificial) systems. Toward this end, I attempt to address consciousness and autonomy on functional, algorithmic, and implementational levels of analysis (Marr, 1983). Finally, I discuss implications of theories of consciousness for the enduring problems of artificial intelligence.

The Enduring Problems of Consciousness

How could there be "*something that it is like*" to be a physical system (Nagel, 1974; Lycan, 1996)? In introducing the *Hard problem*, Chalmers (1997) contrasted this question with the "easy problem" of understanding how biological processes contribute to different psychological phenomena. Proponents of the Hard problem argue that we could have a complete cognitive science, and yet still not understand consciousness. Could cognition take place "in the dark" without generating any subjective experiences, or *qualia*? Could such philosophical zombies perform all the computations enabled by brains, yet lack subjectivity?

Intellectual positions on these matters range from the more inflationary claim that consciousness is a fundamental aspect of the universe, to the more deflationary claim that the Hard problem will be (dis-)solved by answering the easy problems of cognitive science (Dennett, 2018), with no "explanatory gap" remaining. Others have suggested that these metaphysical questions distract from the more productive endeavor of studying why particular experiences are associated with particular physical processes: i.e., the "real problem" of consciousness (Seth, 2016). Even disagreement about the

generation of the Hard problem has become a topic of philosophical inquiry and has been named the "*meta-problem*" (Chalmers, 2018).

While numerous models have been suggested, none are generally considered to have solved the enduring problems of consciousness. Such a solution would require explanation spanning implementational, algorithmic, and functional levels of analysis, with rich connections to experience. Here, I suggest that this multi-level understanding can be obtained by using FEP-AI to ground and combine leading models of consciousness into a unified framework centered on integrated world modeling (IWMT). This article focuses on IIT and GNWT, and in forthcoming work, I will extend this synthesis to additional models—e.g., Higher-Order Thought theories (Brown et al., 2019; Graziano, 2019; Shea and Frith, 2019)—each of which emphasizes different aspects of the nature(s) of consciousness.

Yet another enduring problem can be found in that there is no clearly agreed upon definition of consciousness. Some theories focus on consciousness as phenomenal experience. Others emphasize consciousness as awareness of knowledge, or "access" (Block, 2008). IWMT's primary focus is explaining means by which biological systems may generate phenomenality, or experience as a subjective point of view (Williford et al., 2018; Feiten, 2020). However, IWMT suggests that a variety of higher-order and meta-cognitive capacities may be required in order to obtain coherent subjectivity—although not necessarily involving either access or explicit self-consciousness (Milliere and Metzinger, 2020)—and thereby an experienced world. More specifically, IWMT's primary claims are as follows:

1. Basic phenomenal consciousness is what it is like to be the functioning of a probabilistic generative model for the sensorium of an embodied–embedded agent.
2. Higher order and access consciousness are made possible when this information can be integrated into a world model with spatial, temporal, and causal coherence. Here, coherence is broadly understood as sufficient consistency to enable functional closure and semiotics/sense-making (Joslyn, 2000; Pattee, 2001; Ziporyn, 2004; Gazzaniga, 2018; Chang et al., 2019). That is, for there to be the experience of a world, the things that constitute that world must be able to be situated and contrasted with other things in some kind of space, with relative changes constituting time, and with regularities of change constituting cause. These may also be preconditions for basic phenomenality (#1), especially if consciousness (as subjectivity) requires an experiencing subject with a point of view on the world.
3. Conscious access—and possibly phenomenal consciousness—likely requires generative processes capable of counterfactual modeling (Friston, 2018; Pearl and Mackenzie, 2018; Kanai et al., 2019; Corcoran et al., 2020) with respect to selfhood and self-generated actions.

In what follows, I attempt to justify these claims by integrating across leading theories of emergent causation and consciousness. This approach draws on the explanatory breadth and embodied cybernetic grounding of the FEP-AI, the focus on irreducible integrative complexity provided by IIT, and the functional and mechanistic details provided by GNWT. IWMT tries to make inroads into the enduring problems of consciousness by synergistically combining the relative strengths (and diverse perspectives) of these theories (**Table 1**).

IWMT: Combining IIT and GNWT With the FEP-AI

This section provides an introduction to FEP-AI, IIT, and GNWT, as well as an initial account of how they may be combined within IWMT. Further details regarding FEP-AI and IIT are explored in subsequent sections, followed by a further integration with GNWT.

FEP-AI

The Free Energy Principle states that persisting systems must entail predictive models to resist entropic mixing (Friston, 2019). That is, to prevent destruction and maintain their forms, systems must adaptively respond to a variety of events, and so must be able to model these events in some capacity (Conant and Ashby, 1970). Beginning from this fundamental principle of nature (Hohwy, 2020), the FEP and Active Inference (FEP-AI) framework (Friston et al., 2017a) proscribes means of satisfying this imperative through minimizing prediction-error (or "free energy") with respect to the models by which systems preserve themselves. In contrast to views in which experience emerges from passive sensations, FEP-AI understands perception as taking place within the context of actions, including foraging for information and resolving model uncertainty. Within this framework, both perception and action are understood as kinds of predictions/inferences regarding the means by which prediction-error might be minimized (hence, "active inference").

Hierarchical predictive processing (HPP) offers powerfully explanatory implementational and algorithmic details for active inference (Clark, 2016), providing a single mechanism for both perception and action. FEP-AI further emphasizes the roles of embodiment, selfhood, and agency in minimizing free energy via action–perception cycles, thus naturally supporting bridges to phenomenology on multiple levels. While probabilistic modeling may narrow explanatory gaps between brain and mind, the question remains: how do (seemingly definite) subjective experiences emerge from probabilities?

IIT: Informational Synergy Through Balanced Integration and Differentiation; of MICE and MAPs

IIT begins from phenomenology (Tononi, 2004), observing that consciousness is distinct in its particular details (i.e., information), while also being experienced holistically (i.e., integration). This observation generated the hypothesis that consciousness results from the ability of nervous systems to support diverse state repertoires, while also synergistically integrating this information into wholes greater than the sum of their parts. IIT further suggests that this capacity for generating integrated information can be quantified by analyzing the degree to which systems are irreducible to the information contained in their parts considered separately. IIT has developed through multiple iterations, most recently formalized with phenomenological axioms and the postulated properties required

TABLE 1 | Comparisons between four perspectives on aspects of consciousness: FEP-AI, IIT, GNWT, and IWMT.

	FEP-AI	IIT	GNWT	IWMT
Levels of analysis emphasized	Functional, algorithmic, and implementational	Phenomenological and implementational	Functional and implementational	Phenomenological, functional, algorithmic, and implementational
Emphasizes either phenomenal or access consciousness	Both	Phenomenal	Access	Both
Emphasizes either intrinsic or extrinsic perspectives	Both	Intrinsic	Extrinsic	Both
Neural substrates of consciousness	A distributed pattern of effective connectivity (entailing Bayesian beliefs) across a multi-level deep temporal hierarchy, primarily generated by L5 pyramidal neurons and thalamic relays	A maximal nexus of self-cause–effect power, likely centered on posterior cortices	A global workspace realized by re-entrant connectivity between frontal and posterior cortices	Agreement with FEP-AI, except these distributed patterns are hypothesized to be integrated via the formation of self-organizing harmonic modes, so promoting communication through coherence Agreement with IIT with respect to basic phenomenal consciousness, but with specific emphasis on posterior-medial cortices as a basis for egocentric perspective Agreement with GNWT with respect to access consciousness, but with phenomenality being generated from posterior loci
Minimally conscious system	Any generative model with temporal depth and counterfactual richness; e.g., all deep belief hierarchies capable of adaptive active inference	Any system capable of generating irreducible cause–effect power over itself; e.g., a single elementary particle	Any system capable of implementing a global workspace; e.g., a computer program with a blackboard architecture	Any process capable of generating a world model with spatial, temporal, and causal coherence with respect to the system and its causal inter-relations with its environment; e.g., all mammals, possibly all vertebrates, and possibly insects
Can a system without dynamics be conscious?	No	Yes, if it is part of a configuration capable of constraining likely past and future states	No	No
Could an artificial intelligence (AI) implemented on a von Neumann architecture be conscious?	Yes	No	Yes	Probably
Is either physical or a richly structured virtual embodiment required for consciousness?	Yes	No	No	Yes
Associated concepts from machine learning and AI	Variational autoencoders Forney factor graphs with marginal message passing	Direct implementation on neuromorphic hardware capable of recurrent dynamics	Blackboard architectures	Folded variational autoencoders with recurrent dynamics Turbo codes
Are human-equivalent intelligent zombies feasible?	No comment	Yes	No comment	Theoretically conceivable, but practically infeasible

FEP-AI, Free Energy Principle and Active Inference framework; IIT, Integrated Information Theory; GNWT, Global Neuronal Workspace Theory; IWMT, Integrated World Modeling Theory.

for realizing these aspects of experience in physical systems (Tononi et al., 2016). These postulates are stipulated to be not only *necessary*, but also, controversially (Bayne, 2018; Lau and Michel, 2019), *jointly sufficient* conditions for conscious experience (**Table 2**).

IIT is both a theory of consciousness and meta-physical formalism, attempting to answer the question: what counts as a system from an intrinsic perspective (Fallon, 2018)? IIT models systems as networks of causal relations, evaluating compositional structures for their ability to inform (or constrain) past and future states. *Integrated information (phi)* is calculated based on the degree to which cutting systems along a *minimum information partition (MIP)* impact past and future self-information, evaluated across all relevant spatial and temporal grains for system evolution. The extent to which MIPs reduce self-information is used to calculate the degree to which systems make

TABLE 2 | Integrated Information Theory (IIT) axioms and postulates, with corresponding examples of experiences and mechanistic systems.

IIT axioms: Essential properties of experience	**Example experiences**	**IIT postulates: Properties of physical systems capable of accounting for experience**	**Example systems**
Intrinsic existence:			
Experience exists from its own intrinsic perspective (i.e., subjectivity), independent of external observers.	My experience of a red apple has intrinsic existence in that it is both real to me and also private.	A system has cause–effect power upon itself; present states must inform probabilities of past and future states, so linking causes and effects.	A brain has internal connectivity that influences which states are likely to flow from the past to the future, given its present state; some parts of brains have more intrinsic connectivity than others.
Composition:			
Experience is structured by the elementary or higher-order subjective distinctions out of which it is composed.	My experience of a red apple is composed of particular features, such as redness for color and apple shape for form.	A system is structured by the more elementary sub-systems out of which it is composed, and which have cause–effect power upon the system.	A brain is composed of neurons, whose particular configurations influence its past and future states; different parts of brains have different compositions.
Information:			
Experience is particular in being composed of a specific set of subjective distinctions, so being differentiated from other possible experiences.	My experience of a red apple is informative in being perceived in terms of particular qualities of subjective redness (as opposed to greenness) and apple shape (as opposed to pear shape).	A system specifies a particular cause–effect structure that informs particular probabilistic repertoires of past causes and future effects for the system and sub-systems, so differentiating particular states from other possible states.	A brain can be configured in many different ways, and so any particular configuration is highly informative in terms of being distinguished from other possible configurations; some parts of brains are more informative than others in different contexts.
Integration:			
Each experience is unified in being irreducible to independent subsets of subjective distinctions.	My experience of a red apple is integrated in that redness and apple shape are not independently perceived, but are instead combined into a unified whole.	A system specifies a unified cause–effect structure that is irreducible to independent sub-systems (phi > 0), including its minimally interdependent component.	A brain has properties that do not exist in its individual neurons considered separately; some parts of brains are more integrated than others in different contexts.
Exclusion:			
Each experience is definite in content and spatiotemporal grain, specifying a particular set of subjective distinctions unfolding on particular spatiotemporal scales.	My experience of a red apple has particular contents with respect to space and time, with particular redness and apple shape being perceived at some spatiotemporal scales and not others.	A system specifies particular cause-effect repertoires over particular sets of elements at particular spatial and temporal grains. The boundaries of a system are defined by a complex entailing a maximally irreducible conceptual structure (MICS) existing at particular spatial and temporal grains, whose total integrated information is quantified as Phi-max.	A brain and its sub-systems have particular boundaries that determine the extent to which they function as integrative wholes in space and time; some parts of brains have clearer boundaries than others in different contexts (e.g., modularity).

irreducible (i.e., integrated) causal differences to themselves, thus defining their integrated information (quantified as phi). Intuitively, if something can be decomposed into parts without consequence, then it is not an integrated system. According to the exclusion axiom, systems are only real (and potentially conscious) if they represent maxima of integrated information. The self-directed causal relation of a maximal complex is referred to as a *maximally irreducible conceptual structure (MICS)*—corresponding to mappings onto an abstract metric space (i.e., "qualia space") (Balduzzi and Tononi, 2009), whose particular geometries correspond to particular experiences. Further, sub-mechanisms contributing given MICS will be associated with a variety of phenomenal distinctions, specified as *maximally irreducible cause-effect (MICE) repertories*.

While IIT's experience-first approach provides compelling bridges between phenomenology and mechanistic implementations, the question remains: why should there be "anything that it is like" to be a maximally irreducible cause-effect structure? As described below, IWMT proposes that a maximal complex (entailing a MICS) could also entail subjective experience, *if (and only if)* these complexes also entail probabilistic mappings—or maximal a posteriori (MAP) estimates derived thereof—entailed by generative models for the sensoriums of embodied–embedded goal-seeking agents. As described in further detail below, IWMT further proposes that phi parameterizes the ability of systems to minimize free energy and maximize self-model evidence. While the most valid means of defining integrated information for conscious (or unconscious) systems remains contested (Barrett and Mediano, 2019), one potential advance from IWMT's proposed synthesis could be identifying the appropriate uses for various formulations of integrative complexity.

The putative sufficiency of IIT's phenomenological postulates for consciousness results in a surprising implication: the degree

to which systems exist is also the degree to which they generate conscious experience (Tononi and Koch, 2015). As will be described in greater detail below, IWMT accepts a modified version of this proposition with fewer protopansychist implications: systems exist to the degree they generate model evidence for themselves, which may entail consciousness if models have spatial, temporal, and causal coherence for systems and world. Below, I describe how systems might be configured if they are to generate complexes of integrated information with these coherence-enabling properties.

[Note: A more detailed discussion of IIT's postulates and axioms can be found in IWMT Revisited (Safron, 2019a), in the section: "A review of IIT terminology."]

GNWT: Functional Synergy Through Balancing Integrated and Segregated Processing; Critical Modes of Consciousness as Bayesian Model Selection

Originally introduced by Baars (1993), Global Workspace Theory considers computational requirements for intelligent functioning, drawing analogies between consciousness and computing architectures in which "blackboards" share information among multiple specialist processes. According to Baars, consciousness is hypothesized to correspond to a "global workspace" that allows unconscious segregated processes to communicate with informational synergy. Information becomes conscious when it enters workspaces, and so can be effectively broadcast throughout entire systems. Because of workspaces' limited capacities, specialist processes compete and cooperate for selection based on abilities to satisfy context-specific computational objectives. Workspace architectures have been used in artificial intelligence (Hofstadter and Mitchell, 1994; Shanahan and Baars, 2005; Madl et al., 2011) because of their capacity for integrative functioning with competition-enhanced efficiency. These systems have also been configured in ways that recapitulate notable psychological phenomena, including cognitive cycles involving separable phases of sensing, interpreting, and acting.

The ability of workspaces to "select" value-enhancing information was interpreted as instantiating a quasi-Darwinian process by Edelman et al. (2011). According to neural Darwinism, the functionality of global workspaces provides a computational-level description of a mechanistic "dynamic core," which promotes activity for particular neuronal ensembles through re-entrant connectivity. In line with theories emphasizing binding via synchronous dynamics (Singer, 2001; Varela et al., 2001; Crick and Koch, 2003), the thalamocortical system has been suggested to play key roles in this value-dependent selection and broadcasting of neuronal information.

In terms of neuronal architecture, van den Heuvel and Sporns (2011) have identified connectomic "rich club" networks, whose high centrality and interconnectivity may allow systems with mostly local connections to achieve both integrated and differentiated processing (Sporns, 2013). Shanahan (2012) has further noted that these core networks may be related to intelligence—and presumably consciousness—in non-human animals. Intriguingly, with respect to global workspaces, varying degrees of functional connectivity between richly connected networks have been found to be accompanied by periods of either high or low modularity (Betzel et al., 2016), consistent with a potential functional significance of integrating information across otherwise isolated sub-systems. More recent work (Esfahlani et al., 2020) has demonstrated that transient periods of strong co-activation within these networks explains much of the overall variance and modularity with respect to network structures, consistent with alternating periods of integration and segregation via workspace dynamics.

Within this paradigm of consciousness as enabling the integration and broadcasting of information, Dehaene (2014) has made invaluable contributions in describing how biological implementations of workspace dynamics may help to explain otherwise mysterious aspects of cognition (e.g., psychological refractory periods, attentional blinks). Dehaene et al. have also characterized time courses for unconscious and conscious information processing, showing how transitions to conscious awareness correspond to non-linear increases in large-scale brain activity. These "ignition" events are stipulated to indicate the accumulation of a critical mass of mutually consistent information—implemented by converging excitatory neural activity—so selecting one interpretation out of multiple possibilities. This neurobiological account in which neuronal systems dynamically move between more integrated and segregated processing is referred to by Dehaene and Changeux (2005) as GNWT. From an FEP-AI (and IWMT) perspective, these phase transitions may correspond to discrete updating and Bayesian model selection with respect to perception and action (Friston et al., 2012a; Hohwy, 2012; Parr and Friston, 2018b). GNWT has been increasingly described in terms of Bayesian inference (Dehaene, 2020; Mashour et al., 2020), including in a recently proposed Predictive Global Neuronal Workspace model (Whyte and Smith, 2020).

If neural dynamics can select particular interpretations of events, formally understood as Bayesian inference, then we seem even closer to closing explanatory gaps between mind and brain. Yet, the enduring problems of consciousness remain: Why should it be (or "feel") like something to be a probabilistic model, and which biophysical processes specifically enable workspace-like dynamics?

FEP-AI + IIT + GNWT = IWMT

IIT focuses on consciousness as emerging from systems that are both unified and differentiated through their internal cause–effect relations. GNWT focuses on consciousness as emerging from systems that allow both global and local processing to be balanced through cycles of selecting, amplifying, and broadcasting information. In these ways, IIT and GNWT have identified highly similar preconditions for subjective experience.

While there are extensive similarities between GNWT and IIT, there are also notable differences (**Table 1**). GNWT focuses on systems engaging in cognitive cycles of acting and perceiving. This focus on integrative agentic functioning is highly compatible with the enactive bases of FEP-AI, where action–perception cycles are driven by rounds of Bayesian model selection. IIT

has a broader scope, ascribing consciousness to all systems self-governed by emergent causes. As discussed below, this suggestion is partially compatible with FEP-AI, albeit with a restricted interpretation of the meanings of integrated information as potentially being necessary, but not sufficient for consciousness (Lau and Michel, 2019).

With respect to the neural substrates of consciousness, IIT identifies a "posterior hot zone" (Boly et al., 2017), which has been stipulated to represent a maximum of phi in the brain (Boly et al., 2017), and potentially also a source of spatial phenomenology, due to its organization as a hierarchy of 2D grids (Haun and Tononi, 2019). [Note: This stipulation is currently purely theoretical, as the computations required to formally identify maximal complexes are intractable for biological systems, and it remains contested which estimation methods are most valid in which contexts (Mediano et al., 2019b).] GNWT, in contrast, suggests that consciousness and global availability are made possible by connectivity between posterior and frontal regions. IWMT considers both positions to be accurate, but with respect to basic phenomenal and access consciousness, respectively.

Some of this dispute regarding the neural substrates of consciousness could potentially be resolved by identifying multiple types of workspace (and integrating) dynamics. One way of achieving widespread availability may be via synchronous stabilization (Humphrey, 2017) of representations, or as I suggest below, via *self-organizing harmonic modes (SOHMs)*. These processes may center on posterior hot zones, with information taking the form of a distributed causal nexus with both intrinsic integrated information and extrinsic functional significance. Alternatively, availability may also be achieved via the re-representation and accessing of information. These processes may also center on posterior (particularly medial) cortices as substrates for abstract (low-dimensional) features, potentially providing the kinds of representations adduced by symbolic cognitive science. Global availability and meta-awareness for this information would depend on coupling with the frontal lobes—which would also provide goal-oriented shaping of dynamics—although phi maxima and experience itself might still be generated in posterior hot zones as loci for embodied simulation (Barsalou, 2008, 2009, 2010; Prinz, 2017).

[Note: More details regarding neural substrates of consciousness are described below, as well as in IWMT Revisited (Safron, 2019a) *in the sections: "Neural systems for coherent world modeling" and "Future directions for IIT and GWT."]*

Selfhood, Autonomy, and Consciousness

By grounding IIT and GNWT within the body-centered perspective of FEP-AI, IWMT suggests that complexes of integrated information and global workspaces can entail conscious experiences *if (and only if)* they are capable of generating integrative world models with spatial, temporal, and causal coherence. These ways of categorizing experience are increasingly recognized as constituting essential "core knowledge" at the foundation of cognitive development (Spelke and Kinzler, 2007). In addition to space, time, and cause, IWMT adds embodied autonomous selfhood as a precondition for integrated world modeling. As suggested by Kant (1781) (cf. transcendental unity of apperception), Helmholtz (1878), Friston (2017), and others—e.g., von Uexküll (1957), Damasio (2012), and Humphrey (2017)—IWMT argues that integrated selfhood and autonomy are required for coherent sense-making. For there to be "something that it is like"—and even more so, "something it feels like"—workspace dynamics must be grounded in models of autonomous embodiment (Safron, 2019a,c).

With respect to autonomy, IWMT further suggests that driving of cognitive cycles by "ignition" events may be an apt description. That is, if workspace dynamics implement Bayesian model selection—driven by the minimization of free energy—then cognitive cycles may be fully isomorphic with both thermodynamic work cycles (Kauffman and Clayton, 2006; Deacon, 2011) and selective pressures in the context of generalized Darwinism (Kaila and Annila, 2008; Campbell, 2016; Safron, 2019b). That is, if ignition corresponds to large-scale updating and communication of Bayesian beliefs, then formally speaking, these events may be sources of cause–effect power in precisely the same ways that controlled explosions drive engines to generate work. If these beliefs entail intentions for acting and the phenomenology of willing, then will power may be a systemic cause and source of force in every meaningful sense of the words "power," "cause," and "force" (Carroll, 2016; Sengupta et al., 2016; Pearl and Mackenzie, 2018; Safron, 2019c; Friston et al., 2020b).

As described below, this connection to autonomy is yet another way in which IIT and GNWT may be synergistically combined: the ability of workspaces to support cognitive cycles may depend on maintaining coherent internal dynamics, which may also depend on exerting cause–effect power over themselves. With respect to IIT, maximally irreducible cause-effect structures (MICS) may correspond to maximally probable inferences over sensorimotor states for integrated systems, as well as sources of maximal control energy governing system evolution. Thus, IWMT's cybernetic (Seth, 2015; Safron, 2019c) grounding of IIT and GNWT within FEP-AI may not only help explain why there may be "something that it is like" to be a maximal complex (entailing a MICS and MICE repertoires), but also provide causal connections between consciousness and action, thus providing foundations for the emergence of agency (Tononi, 2013).

The *default mode network (DMN)* and functional networks with which it interacts (Huang et al., 2020) may be particularly important for understanding the emergence of both phenomenal and higher-order consciousness, and also agency. In predictive processing, intentional action selection requires an ability to maintain counterfactual predictions in the face of otherwise inconsistent sense data (Safron, 2019c). However, driving systems into otherwise uncharted territories of inference-space will involve temporary local increases in prediction-error (i.e., "free energy") for portions of generative models that recognize discrepancies between imagined goal states and current sensory observations. In order for goal-oriented behavior to proceed, this free energy must be buffered by other systems capable of acting as temporary thermodynamic reservoirs (Carhart-Harris and Friston, 2010). The DMN and its imaginative capacities (Beaty et al., 2014, 2015, 2018; Hassabis et al., 2014) may instantiate this kind of (informational) creative dynamo, constituting sources

of strongly internally coherent predictions, thus being capable of temporarily absorbing and then releasing free energy via the shaping of perception and driving of action. The network properties of the DMN are ideally suited to serve these functions, having both high centrality—and so high potential for integrating information and exerting control (Kenett et al., 2018)—while also being located distally from primary modalities, and so being capable of supporting dynamics that are more decoupled from immediate sensorimotor engagements (Sormaz et al., 2018; Corcoran et al., 2020). Further, the DMN is likely to support some of the most stable inferences available to embodied–embedded persons, with major nodes allowing for egocentric perspective—i.e., providing a subjective point of view in generating world models with spatial, temporal, and causal coherence—integrated memory, and even the foundations of selfhood (Dennett, 1992; Hassabis and Maguire, 2009; Northoff, 2012; Brewer et al., 2013; Davey and Harrison, 2018). Indeed, the DMN and the networks with which it couples may be well-modeled as a complex of effective connectivity with high degrees of integrated information, functioning as a dynamic core and global workspace for conscious imaginings (Wens et al., 2019). In these ways, and as will be described in greater detail below, IWMT suggests that a multi-level account of the nature of embodied experience and its connections to phenomenology may contribute to the quest for obtaining satisfying solutions to the Hard problem.

FEP-AI AND IIT: UNIFIED SYSTEMS THEORIES

The following sections discuss FEP-AI and why it is increasingly recognized as a unified systems theory. I will also suggest ways that IIT can be integrated with FEP-AI, thereby illuminating the nature of consciousness and causal emergence more generally. Readers specifically interested in the neurocomputational bases of consciousness may want to skip to "Mechanisms of Integrated World Modeling." However, this is not recommended, as earlier sections help to show how FEP-AI provides a multi-level grounding for other theories in fundamental biophysics, thus linking mind and life. These sections also help to clarify what is and is not implied by these frameworks (i.e., which systems are likely to have or lack consciousness), as well as the implications of their integration for understanding emergent complexity in multiple domains.

Resisting the 2nd Law With Generative Modeling (and Integrated Information)

According to the 2nd law, systems should exhibit increasing disorder until they cease to exist. Yet some things do manage to (temporarily) persist, and so something about their configurations must organize environmental exchanges to avoid entropic accumulation (Schrodinger, 1944; Brillouin, 1951; Deacon, 2011; Ramstead et al., 2018). Persisting systems somehow generate dynamics that steer away from the maximally probable outcome of maximal disorder. In cybernetics and control theory, the requirements for such governing processes are expressed as the *good regulator theorem* and *law of requisite variety*: any effective controller must be able to (at least implicitly) model that system, and regulating models require sufficient complexity to represent the variety of states likely to be encountered (Conant and Ashby, 1970).

FEP-AI (Friston, 2019) views persisting systems as entailing generative models for the preconditions by which they persist. For a system to constitute a model, its composition must be able to either compress or predict information for that which is modeled. Persisting systems specifically generate mutual (probabilistic) information between past and future states based on their present compositions. These mappings between particular configurations and ensuing dynamics constitute likelihoods (as particular action tendencies), thus characterizing system compositions as generative models, which generate dynamics that maximize the probability of those particular compositions. If it were not the case that system configurations generate dynamics that maintain those configurations, then no persisting systems would exist. Thus, persisting systems can be viewed as generative models that generate evidence for themselves through their dynamics, and so engage in "*self-evidencing*" (Hohwy, 2016).

In this way, FEP-AI provides a formalization and generalization of autopoietic self-making as described by Maturana and Varela (1980):

> *"An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in space in which they (the components) exist by specifying the topological domain of its realization as such a network."*

To the degree systems persist, they possess attracting sets that define them as particular phase space densities—whose action constitutes trajectories through state space—with varying probabilities of occurrence. In autopoiesis, attractor dynamics produce the very mechanisms out of which they are generated. FEP-AI views these autopoietic attractor configurations and ensuing trajectories as self-predicting generative models (Palacios et al., 2020), where that which is generated is the very probabilistic densities that define the existence of particular systems.

FEP-AI goes on to quantify self-model evidence according to an information-theoretic functional of variational (or approximate) free energy (Dayan et al., 1995). Derived from statistical physics, this singular objective function is optimized by minimizing discrepancies between probabilistic beliefs and observations (i.e., prediction-error, or "surprisal"), penalized by model complexity. To the extent systems persist, they constitute existence proofs (Friston, 2018) that they were able to bound surprise (i.e., high-entropy configurations) relative to predictive models by which they perpetuate themselves. Systems must

respond adaptively to a variety of situations in order to avoid entropy-increasing events, and so must entail models with sufficient complexity to predict likely outcomes, thus minimizing discrepancies between expectations and observations. However, these models must not have so much complexity that they waste energy or over-fit observations and fail to generalize their predictions (also, more complex models are more energetically costly to implement). Variational free energy provides an objective function that optimally balances these requirements for accuracy and simplicity.

The extreme generality of FEP-AI requires emphasis. Not only do nervous systems entail predictive models, but so do entire populations of organisms and their extended phenotypes (Dawkins, 1999) as teleonomical (Deacon, 2011; Dennett, 2017) predictions with respect to evolutionary fitness (Friston, 2018; Ramstead et al., 2018). By this account, nervous systems are merely a (very) special case of generative modeling, where *all systems are models* in their very existence, but where *some systems also have sub-models* that function as cybernetic controllers (Stepp and Turvey, 2010; Seth, 2015; Seth and Tsakiris, 2018). In these ways, FEP-AI provides a formalism where persisting dynamical systems can be understood as self-generating models, grounded in first principles regarding the necessary preconditions for existence in a world governed by the 2nd law.

This view of systems as self-predicting generative models has clear correspondences with IIT, since self-evidencing depends on capacity for generating self-cause–effect power. I suggest we should further expect model-evidence for system preservation to be related to a system's ability to function as a unified whole, and so integrated information maximization ought to accompany free energy minimization. Notably, IIT-based models of metabolic cycles and gene-regulatory networks—core processes for homeostasis and autopoiesis—suggest that adaptive capacities of biotic systems may require high-phi configurations (Marshall et al., 2017; Abrego and Zaikin, 2019). Systems with lower phi may be qualitatively different from systems with higher phi (Albantakis, 2017; Albantakis and Tononi, 2019), being less capable of state-dependent adaptation—and thereby learning—which may drastically limit their intelligence and agency. These IIT-informed studies are fully consistent with FEP-AI, wherein all persisting systems minimize free energy, but only evolved systems minimize expected free energy via generative models where causes can be modeled with temporal depth and counterfactual richness (Kirchhoff et al., 2018).

An Ontology of Markov Blankets: Estimating Boundaries (and Intelligence-Potential) for Processes/Things as Self-Predicting Models and Complexes of Integrated Information

This formalization of autopoietic systems can also be derived with graphical modeling concepts, providing further bridges between FEP-AI and IIT. Graphical models represent systems as structured relationships among component variables and their connections. If these connected variables are associated with probabilities—whether due to uncertain observations or inherent stochasticity—then that representation is a probabilistic graphical model (PGM) (Koller and Friedman, 2009). PGMs specify probability distributions over variables, thus entailing probabilistic models of that which is represented. This mapping from connected graphs to probabilities allows PGMs to synergistically combine information from multiple sources. Integration into joint probability distributions affords inference of both likely beliefs from observations (i.e., discriminative models) and likely observations from beliefs (i.e., generative models). With importance for subsequent discussions of consciousness, these graphs not only enable the generation of probabilistic world models (i.e., inference) and refinements of these models with observations (i.e., learning), but PGMs also afford discrete estimates of the most likely values for variable combinations, as in maximum a posteriori (MAP) estimation.

For any PGM component, the set of surrounding nodes is referred to as a *Markov blanket (MB)* (Pearl, 1988), which establishes conditional independence between internal and external variables. All paths connecting internal and external states are mediated by MBs; thus, conditioning upon this blanketing set integrates all mutual information across this partition (i.e., marginalization). System MBs define epistemic relationships with the external world in providing the only source of information that internal states ever receive (Hohwy, 2017). Everything beyond MB boundaries is not directly observable, and so latent values of external states must be inferred.

Described as PGMs, the functional boundaries of systems are MBs (Kirchhoff et al., 2018), mediating all that can ever be known about or done to the outside world. Some examples: single-celled organism MBs are largely co-extensive with cellular membranes; nervous system MBs are composed of sensor and effector neurons by which they receive information from sensors and drive change with actuators; niche-constructing organism MBs constitute the boundaries of extended phenotypes, including bodies and external structures that regulate environmental interaction. Such functional boundaries are an essential source of adaptive constraints for biological systems (Rudrauf et al., 2003; Hordijk and Steel, 2015; Lane, 2016), both internally concentrating system-promoting complexity and limiting system-threatening exchanges with external environments. Thus, MBs are both epistemic and system-defining boundaries. With respect to IIT, the boundaries of maximal complexes (entailing maximally irreducible cause-effect structures) would also constitute MBs. Although each MICS represents a kind of world unto itself (Leibniz, 1714), FEP-AI's formalism of internal states as modeling external states (and vice versa) may provide a means of understanding how such inwardly directed phenomena can nonetheless come to "encode" meaningful information about the external world with which they co-evolve, potentially providing linkages between IIT's intrinsic integrated information and information theory more generally.

The dual epistemic and ontological roles of MBs help justify the extremely broad scope of both FEP-AI (and possibly IIT as well). Identifiable systems must have boundaries defining

their extents relative to other systems. Persisting systems further require predictive models to maintain themselves and their MB boundaries as they interact with environments. Yet, because blanket states informationally shield internal states from the rest of the world, modeling external states and MB boundaries necessitates inference (Friston, 2017, 2018, 2019). In this way, the epistemic boundaries created by system-defining MBs require persisting dynamical systems to entail self-evidencing generative models.

Generative Modeling, Integrated Information, and Consciousness: Here, There, but Not Everywhere?

The extreme generality of PGMs and the implicit modeling relationships prescribed by FEP-AI may be of an extremely simple variety, particularly if systems have limited dynamic character and restricted thermodynamic openness. To provide an intuition-stretching example, by virtue of persisting (and so generating model evidence for their existence), the configuration of rocks and resultant causal interactions could be viewed as instantiating an implicit "prediction" that intramolecular forces and limited exchanges will be sufficient to maintain their forms. On short timescales, rocks will be able to (non-adaptively) generate rock-like dynamics, which restrict thermodynamic exchanges, thus allowing rocks to temporarily avoid disintegration. However, in contrast to living systems, rocks lack functional closure (Joslyn, 2000; Pattee, 2001; Deacon, 2011; Gazzaniga, 2018) with the geological processes generating their forms. Without multi-level evolutionary optimization (Safron, 2019b), generative models will be of such simple varieties that they are incapable of predicting and responding to particular events (i.e., adaptation). In this way, rocks are "surprised" by every exchange with their environments capable of altering their structures, and so will steadily disintegrate as such exchanges accumulate over time. [Note: FEP-AI focuses on weakly mixing ergodic systems, and as such, this conceptual analysis of rocks lacks the kinds of formal treatments that have been—controversially (Biehl et al., 2020; Friston et al., 2020a)—applied to complex adaptive systems.]

This consideration of rocks as (very) impoverished generative models provides a limit case for understanding what is and is not implied by FEP-AI: every 'thing' can be viewed as having a basic kind of intelligence by virtue of existing at all, but neither rocks nor other similar inanimate objects are conscious (Friston, 2018, 2019). This limit case also shows major points of intersection between FEP-AI and IIT (**Table 1**), as both frameworks provide universal ontologies, and so must be applicable to every system, including rocks, and potentially even the processes giving rise to physical forces and their associated particles (Tegmark, 2014). However, according to IIT's exclusion axiom, rocks would not represent actual systems, in that maxima of integrated information would likely be found among separate components, and so neither (intrinsic) existence nor quasi-sentience would be ascribed. While the exclusion axiom may be essential for consciousness, relaxing this postulate in some cases may allow IIT to both (a) be fully compatible with FEP-AI and (b) better function as a general model of emergent causation. That is, for something to be said to exist, it may not be necessary for it to be a maximum of integrated information as irreducible cause–effect power. Rocks do indeed exist—while lacking consciousness—in that they possess emergent properties that are not present in their constituent elements (e.g., the intrinsic property of a boulder being able to maintain its form as it rolls (Bejan, 2016), or its extrinsic properties with respect to anything in the path of a large quickly moving object). Large-scale compositions may not represent maximal complexes, but may nonetheless play important roles with respect to internal functioning and interactions with other systems.

With respect to the exclusion principle, IIT theorists have suggested that advanced artificial intelligences could be unconscious "zombies" if deployed on von Neumann architectures (Tononi and Koch, 2015), which lack irreducible integration due to serial operation. However, alternative interpretations of IIT could extend phi analyses into temporally extended virtual processes, rather than solely focusing on "direct" realization by physical mechanisms. From an FEP-AI perspective, maximally explanatory models for computer programs may correspond to (MB-bounded) functional cycles on the software level. This proposal for updating IIT aligns with a recently-suggested theory of consciousness focusing on spatiotemporal scales at which functional closure is achieved (Chang et al., 2019), thus instantiating emergence and affording coarse-graining over lower levels of analysis. However, both Information Closure Theory and IIT purport that consciousness corresponds to any instance of emergent causation. IWMT, in contrast, argues that consciousness may be "what physics feels like from the inside" (Koch, 2012; Tegmark, 2014), *if (and only if)* physical processes support the generation of integrated system–world models with spatial, temporal, and causal coherence.

Consciousness, Emergence, Integrative Synergy

IWMT suggests that leading theories of consciousness can be synergistically combined within FEP-AI. FEP-AI and IIT both play dual roles in this synthesis, serving as both general systems theories and descriptions of the processes underlying subjective experience. FEP-AI and IIT intersect on multiple levels, with potential for understanding causal emergence on multiple scales. However, the nature of these explanations may vary across domains, including with respect to analytic assumptions. Integrated information may potentially be modeled in different (and differently valuable) ways in different contexts (Tegmark, 2016; Mediano et al., 2019a,b), which may range from the identification of natural kinds, to the nature of life, to perception, and even consciousness (**Figure 1**). Based on these considerations, I propose it may be productive to factorize IIT into two complementary versions:

1. IIT-Consciousness: the original version of the theory.
2. IIT-Emergence: an alternative version of the theory where the exclusion axiom is relaxed.

In both cases, IIT would still correspond to an analysis of systems in terms of their irreducible cause–effect power. However, the relaxation of the exclusion axiom in IIT-Emergence

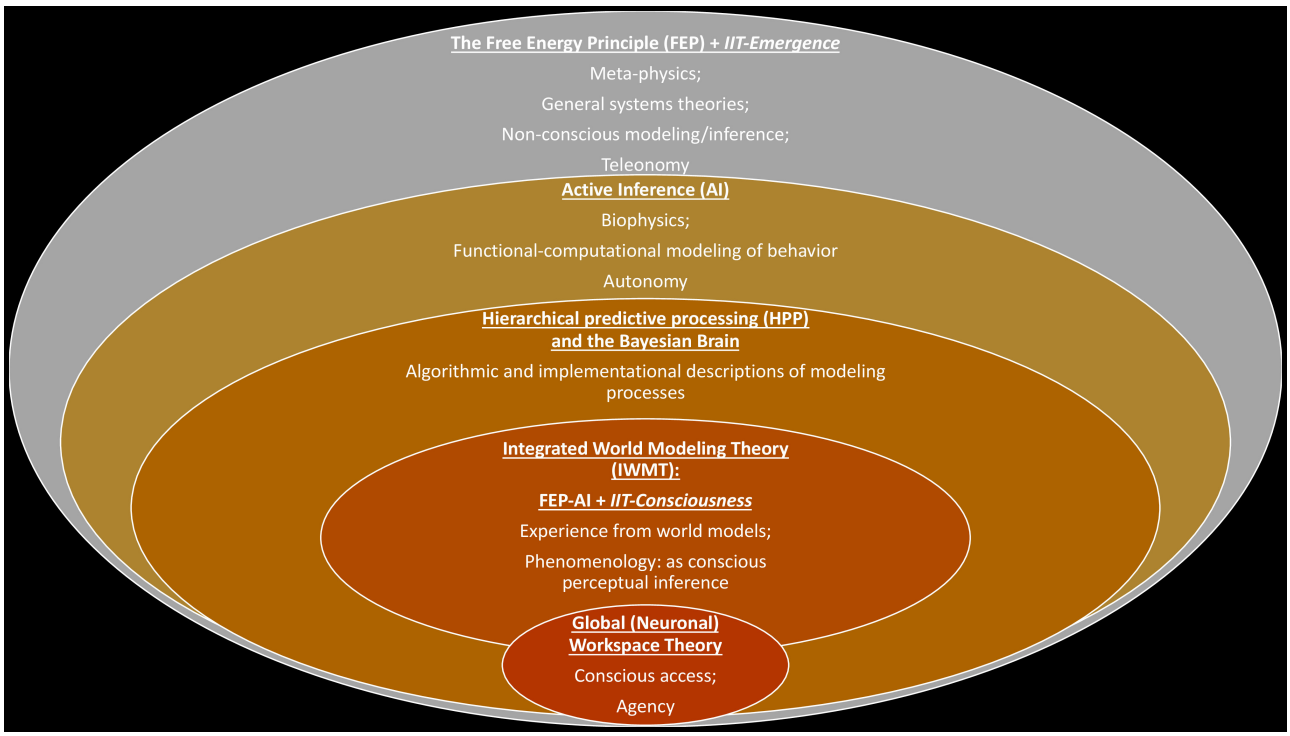


FIGURE 1 | Intersections between FEP-AI, IIT, GNWT, and IWMT.
The *Free Energy Principle (FEP)* constitutes a general means of analyzing systems based on the preconditions for their continued existence via implicit models. *Integrated Information Theory (IIT)* provides another general systems theory, focused on what it means for a system to exist from an intrinsic perspective. The extremely broad scope of FEP-AI and IIT suggests (and requires for the sake of conceptual consistency) substantial opportunities for their integration as models of systems and their emergent properties. Within the FEP (and potentially within the scope of IIT), a normative functional-computational account of these modeling processes is suggested in *Active Inference (AI)*. *Hierarchical predictive processing (HPP)* provides an algorithmic and implementational description of means by which systems may minimize prediction error (i.e., free energy) via Bayesian model selection in accordance with FEP-AI. Particular (potentially consciousness-entailing) implementations of HPP have been suggested that involve multi-level modeling via the kinds of architectures suggested by *Global Neuronal Workspace Theory (GNWT)*. The concentric circles depicted above are intended to express increasingly specific modeling approaches with increasingly restricted scopes. (Note: These nesting relations ought not be over-interpreted, as it could be argued that HPP does not require accepting the claims of FEP-AI.) This kind of generative synthesis may potentially be facilitated by developing an additional version of IIT, specifically optimized for analyzing systems without concern for their conscious status, possibly with modified axioms and postulates: *IIT-Consciousness* (i.e., current theory) and *IIT-Emergence* (e.g., alternative formulations that utilize semi-overlapping conceptual-analytic methods). *Integrated World Modeling Theory (IWMT)* distinguishes between phenomenal consciousness (i.e., subjective experience) and conscious access (i.e., higher-order awareness of the contents of consciousness). Non-overlap between the circle containing GNWT and the circle containing IIT-Consciousness is meant to indicate the conceivability of subjectivity-lacking systems that are nonetheless capable of realizing the functional properties of conscious access via workspace architectures. IWMT is agnostic as to whether such systems are actually realizable, either in principle or in practice.

could afford a more flexible handling of different kinds of emergent causation (e.g., relative cause–effect power from various coupling systems), as well as more thorough integration with FEP-AI. This broader version of IIT could also sidestep issues such as quasi-panpsychism, as integrated information would not necessarily represent a sufficient condition for generating conscious experiences. While this proposal may not resolve all debates between IIT and GNWT, it may provide further opportunities for integration and synergy between these two theories (e.g., applying—not necessarily consciousness-entailing—phi analyses to posterior and frontal cortices during different stages of cognitive cycles).

The Bayesian Brain and Hierarchical Predictive Processing (HPP)

Broadly speaking, nervous systems can be straightforwardly understood as generative probabilistic graphical models (PGMs). The directed structure of neurons and their organization into networks of weighted connections generate patterns of effective connectivity (Friston, 1994), where flows of influence are physical instantiations of conditional probabilities. From this perspective, nervous systems can be viewed as modeling the world to the extent neural dynamics reflect patterns in the world. The *Bayesian brain hypothesis* (Friston, 2010) proposes this mutual information takes the form of probabilistic mappings from observations to likely causes, and that these inferences may approach bounded optimality with respect to ecological decision-theoretic objectives (Russell and Subramanian, 1995; Mark et al., 2010; Hoffman and Singh, 2012) over phylogenetic and ontogenetic timescales.

The Bayesian brain hypothesis is supported by evidence for a common cortical algorithm of *hierarchical predictive processing (HPP)*—a potential Rosetta stone for neuroscience (Mumford, 1991; Rao and Ballard, 1999; Hawkins and Blakeslee, 2004).

In HPP, neuronal processes constitute hierarchically organized generative models, which attempt to predict likely (hierarchically organized) world states that could have caused actual sensory observations (Friston and Kiebel, 2009; Clark, 2013). Bottom-up sensory information is simultaneously predicted across levels by sending predictions—as Bayesian beliefs, or prior expectations—downwards in anticipation of sensory observations. Prediction-errors (i.e., discrepancies with predictions) are passed upwards toward higher levels, whose modifications update beliefs into posterior expectations, which then become new (empirical) predictions to be passed downwards. This coding scheme is Bayesian in implementing the kind of model selection involved in hierarchical hidden Markov models (George and Hawkins, 2009), or hierarchical Kalman filtering. HPP is also Bayesian in that hierarchical updates combine predictions and prediction-errors according to the relative (estimated) precision of these entailed probability distributions, with this precision-weighting constituting an inverse-temperature parameter by which attention is modulated (Friston et al., 2012b). Notably with respect to the present discussion—and as a source of empirical support for HPP—specific functional roles have been proposed for different frequency bands and cell types, with beta and gamma corresponding to respective predictions and prediction-errors from deep and superficial pyramidal neurons (Bastos et al., 2012; Chao et al., 2018; Scheeringa and Fries, 2019). To summarize, in HPP, each level models the level below it, extending down to sensor and effector systems, with all these models being integrated when they are combined into larger (MB-bounded) generative models (e.g., brains and organisms).

Generalized HPP and Universal Bayesianism/Darwinism

Although evidence for HPP is strongest with respect to cortex, efficiency considerations (Harrison, 1952) provide reason to believe that this may be a more general phenomenon. Some evidence for extending HPP to non-cortical systems includes decoding of predictive information from retinal cells (Palmer et al., 2015), and also models of motor control involving spinal reflex arcs as predictions (Adams et al., 2013). HPP may further extend beyond nervous system functioning and into processes such as morphogenesis (Friston et al., 2015)—observed to exhibit near-optimal utilization of information (Krotov et al., 2014; Petkova et al., 2019)—and even phylogeny (Ramstead et al., 2018).

This leads to another surprising implication of FEP-AI: the broad applicability of the MB formalism suggests that *any persisting adaptive system will enact some kind of HPP*. More specifically, MB-bound systems contain MB-bound sub-systems, with nesting relations reflecting levels of hierarchical organization. More encompassing (hierarchically higher) models accumulate information from the sub-models they contain, with relative dynamics unfolding on either longer or shorter timescales, depending on relationships among nested MB-bounded systems. The epistemic boundaries instantiated by MBs mean that internal and external states are latent with respect to one another, and so must be inferred. Therefore, the communication of information regarding sub-system internal states (via MBs, definitionally) to the larger systems of which they are part is the propagation of a probabilistic belief—e.g., marginal message passing (Parr et al., 2019)—and so overall hierarchical organization of systems and sub-systems must instantiate HPP.

This generalized HPP may be supported by the near-ubiquitous phenomenon whereby coupling systems minimize free energy more effectively through forming larger systems via mutual entrainment (Jafri et al., 2016). From an FEP-AI perspective, this coupling relationship is one of mutual modeling and collaborative inference (Friston and Frith, 2015; Friston, 2017; Kirchhoff et al., 2018; Palacios et al., 2019). This generalized synchrony (Strogatz, 2012) has also been characterized in thermodynamic terms (Kachman et al., 2017; Friston, 2019), where systems spontaneously self-organize into resonant modes with the environments with which they couple—i.e., absorb work and minimize free energy according to Hamilton's principle of least action—where coordinated dynamics have been observed to contain mutually predictive information (Friston, 2013). Notably, coupled attractors have recently been found to adjust their dynamics beginning at sparsely frequented areas of phase space (Lahav et al., 2018). If these synchronizing manifolds begin to nucleate from improbable (and so surprising) alignments, this flow of (mutual-information maximizing) influence might be functionally understood as updating via "prediction-errors." While admittedly speculative, these considerations suggest that generalized HPP (and selection for integrated information) could represent a universality class whose potential extensions are nearly as widespread as generalized synchrony itself. Generalized predictive synchrony may also have implications for IIT, potentially helping to explain how internally directed complexes of integrated information can come to resonate with the external world. Further, synchronization dynamics may provide a mechanistic basis for bridging FEP-AI, IIT, and GNWT, as described below with respect to integration via *self-organizing harmonic modes (SOHMs)*.

Free energy may be most effectively minimized—and integrated information maximized (Marshall et al., 2016)—if synchronized couplings take the form of hierarchically organized modules, thus affording robustness, separable optimization, balanced integration and differentiation, evolvability via degeneracy, efficient communication via small-world connectivity, and flexible multi-scale responsivity via critical dynamics (Meunier et al., 2010; Wang et al., 2011; Ódor et al., 2015; Lin and Tegmark, 2017; Lin et al., 2017; Gazzaniga, 2018; Takagi, 2018; Badcock et al., 2019). Hierarchical organization, modularity, and *self-organized criticality (SOC)* may promote both integrated information maximization and free energy minimization (Friston et al., 2012a, 2014; Vázquez-Rodríguez et al., 2017; Hoffmann and Payton, 2018; Salehipour et al., 2018; Khajehabdollahi et al., 2019), potentially suggesting major points of intersection between FEP-AI and IIT across a wide range of systems.

For biological systems, cells integrate information unfolding at cellular scales, with tissues and organs integrating this information at organismic scales, with organisms and groups of organisms integrating this information at even broader scales.

It is important to remember that FEP-AI can be viewed as a Bayesian interpretation of *generalized Darwinism* (Kaila and Annila, 2008; Harper, 2011; Frank, 2012; Campbell, 2016), and so these nested couplings can also be viewed in terms of natural selection and niche construction unfolding over multiple hierarchical scales (Constant et al., 2018; Ramstead et al., 2018; Badcock et al., 2019). More specifically, a hierarchy of MBs constitute a hierarchy of selective pressures (Safron, 2019b), with dynamics on one level being selected by the next level of organization. These informational shielding properties of MBs connect with debates regarding units of selection in evolutionary theory, in that only organismic phenotypes—and sometimes groups of organisms (Laland et al., 2015; Richerson et al., 2016)—are "visible" to natural selection with respect to phylogeny. However, specific phenotypes are determined by interactions between internal intrinsic dynamics (i.e., intra-system evolution) as well as external systems with which these dynamics couple via niche construction and phenotypic plasticity (Constant et al., 2018). To the (necessarily limited) extent these adaptively coupled nested scales are shaped by stable selective pressures, then the transmission of information across levels could approach Bayes-optimal (Kaila and Annila, 2008; Payne and Wagner, 2019) active inference by combining all relevant probabilistic influences via gradient ascent/descent over fitness/energy landscapes. That is, what is actively inferred by systems (as generative models) in FEP-AI is the inclusive fitness of the sum-total of all quasi-replicative (i.e., self-evidencing) dynamics capable of interacting on the spatial and temporal scales over which evolution (as inference) occurs.

While this discussion of Bayesian generalized Darwinism may seem needlessly abstract, this multi-level account is essential for understanding what we ought to expect to be generated by competing and cooperating quasi-replicative processes (i.e., evolution). It also provides another potential point of intersection with IIT, in that some dynamics will be more influential than others on the timescales at which interactions occur. Specifically, when considered as networks of relations, some sub-graphs will have more integrated information (i.e., intrinsic cause–effect power, or phi) than others, and phi associated with these subgraphs may parameterize capacity to shape overall directions of evolution.

Importantly, if evolution (as inference) applies not just on the level of phylogeny, but also to intra-organism dynamics, then this provides a means of understanding mental processes as both Bayesian model selection and a kind of (generalized) natural selection (Edelman, 1987). With respect to IIT, the irreducible internal cause-effect power for a particular subnetwork of effective connectivity may correlate with its degree of external cause-effect power in influencing the overall direction of evolution within a mind. If a subnetwork of effective connectivity entails a generative model for enacting particular (adaptive) system–world configurations, then a maximal complex of integrated information would also be a maximally explanatory model for overall system evolution, which may entail consciousness under certain conditions.

In this way, FEP-AI shows how mental causation may be isomorphic with evolutionary causation (i.e., action selection as generalized natural selection), where selective pressures constitute free energy gradients, thus providing formal connections with thermodynamic pressures and power-generation abilities. Power is force integrated over time, which may be more likely to be generated by systems capable of exerting cause–effect power over themselves, suggesting a potentially important role for integrated information in modeling evolutionary dynamics. In this way, by describing mental processes in terms of degrees of self-directed cause–effect power, IIT may help explain how particular processes—including those entailing beliefs and desires—possess varying capacities for contributing to informational and thermodynamic work cycles (Kauffman and Clayton, 2006; Deacon, 2011). Taken together, FEP-AI and IIT show how consciousness may not only represent a system's best guess of what is happening at any given moment, but a source of maximal control energy for system evolution, thus providing a means by which conscious intentions can have causal powers.

While HPP is an extremely broad framework, the difference between *basic active inference* and *adaptive active inference* is important to remember (Kirchhoff et al., 2018): while FEP-AI views all systems *as* models, only some of these models afford adaptivity, and only some systems also *have* models (Seth and Tsakiris, 2018). Living organisms possess specific sub-systems capable of supporting generative models with *temporal depth* and *counterfactual richness* (Friston et al., 2017c). These sub-systems are called brains, and they allow organisms to navigate exchanges with their environments by modeling not just present world configurations, but also possible world configurations predicted based on future (counterfactual) actions (i.e., expected free energy).

Brains acquire especially powerful predictive modeling abilities when they are organized according to multiple layers of hierarchical depth. This deep organization allows these systems to model not only transient events at lower levels, but also their organization into more temporally extended sequences at higher levels (Hawkins and Blakeslee, 2004; Baldassano et al., 2017; Friston et al., 2017c). Further, deep internal dynamics create a potential for functional decoupling between modeling and the unfolding of particular sensorimotor engagements (Tani, 2016; Sormaz et al., 2018; Corcoran et al., 2020), thus enabling counterfactual simulations (Kanai et al., 2019) with temporal "thickness"/"depth" (Humphrey, 2017; Friston, 2018), which when conscious enable imagination and explicit planning. These capacities afford the possibility of constructing rich causal world models (Hassabis and Maguire, 2009; Buchsbaum et al., 2012; Pearl and Mackenzie, 2018; MacKay, 2019), and as discussed below, preconditions for coherent conscious experience. In this way, while all brains may expand autonomous capacity by engaging in HPP, only some architectures may be capable of supporting flexible cognition. Thus, FEP-AI implies a near universality for generative modeling, but not necessarily for consciousness. We will now explore properties of nervous systems that may be particularly important for enabling conscious experiences via complexes of integrated information and global workspaces.

MECHANISMS OF INTEGRATED WORLD MODELING

Self-Organizing Harmonic Modes

IWMT proposes a mechanism by which complexes of integrated information and global workspaces may emerge as metastable synchronous complexes of effective connectivity, or *self-organizing harmonic modes (SOHMs)*. SOHMs are proposed to be attractors and eigenmodes (Friston et al., 2014)—or, solutions to harmonic functions—for phase space descriptions of system dynamics, with particular boundaries depending on network topologies over which synchronization occurs. This view of dynamical systems in terms of SOHMs can be understood as an extension of Atasoy et al.'s (2018) analytical framework wherein spectral decomposition is used to characterize brain activity as mixtures of "*connectome harmonics*." When this method was first introduced, Atasoy et al. (2016) compellingly demonstrated how reaction-diffusion simulations of spreading activation could generate resting state networks as stable modes—or standing waves—so recapitulating well-known patterns of neuronal organization with minimal assumptions. Intriguingly, hallucinogenic compounds expanded the repertoire of these harmonic modes (Atasoy et al., 2017), increasing spectral diversity and shifting the distribution of modes toward power-law distributions, a putative—albeit controversial (Touboul and Destexhe, 2017)—hallmark of criticality (Fontenele et al., 2019). This finding is consistent with other studies of psychedelic compounds (Tagliazucchi et al., 2014; Schartner et al., 2017; Viol et al., 2017), supporting the hypothesis that brains may enhance dynamical reconfigurability by being "tuned" toward near-critical regimes (Pletzer et al., 2010; Haimovici et al., 2013; Carhart-Harris, 2018).

Atasoy et al. (2016) describe this modeling approach of identifying eigenfunctions (over a system's Laplacian) as having an extremely broad scope, with applications ranging from Turing's (1952) account of morphogenesis, to acoustic phenomena and other patterns observed with vibrating media (Ullmann, 2007), to solutions for electron orbitals in quantum mechanics (Schrödinger, 1926). Based on our previous discussion of probabilistic graphical models as a near-universal representational framework, the term "connectome harmonics" could be reasonably generalized to apply to all systems. However, IWMT introduces the new term of "SOHMs" to prevent confusion and to emphasize the dynamic self-organizing processes by which synchronous complexes may emerge, even when constituting local standing wave descriptions over dynamics (rather than constituting a Fourier basis for an entire connected system). That is, Atasoy's connectome harmonics constitute a more specific—and important for the sake of understanding consciousness—variety of SOHM.

SOHMs may act as systemic causes in selecting specific dynamics through synchronous signal amplification, with micro-dynamics having greater contributions to synchronizing macro-dynamics when phase-aligned. SOHMs could be viewed as either standing or traveling waves, depending on the level of granularity with which they are modeled (Friston et al., 2014; Mišić et al., 2015; Atasoy et al., 2018; Muller et al., 2018; Zhang et al., 2018). However, when viewed as harmonic modes, SOHMs would have specific boundaries and timescales of formation. In this way, resonant signal amplification within SOHMs could select patterns of effective connectivity based on the timescales at which maximal coherence is achieved. IWMT specifically proposes that these synchronous complexes promote "*communication through coherence*" (Hebb, 1949; Dehaene, 2014; Fries, 2015; Deco and Kringelbach, 2016; Hahn et al., 2019). From an FEP-AI perspective, this synchrony-enhanced communication would facilitate information sharing among (and marginalization over) coupled dynamics, thereby organizing message passing (or belief propagation) for inference (Parr and Friston, 2018a; Parr et al., 2019).

With respect to emergent causation, *circular causal processes* by which SOHMs form would constitute organization into renormalization groups and attracting flow paths along center manifolds (Haken, 1977, 1992; Bogolyubov and Shirkov, 1980; Li and Wang, 2018; Shine et al., 2019). This synchronization of micro-scale phenomena into larger groupings on meso- and macro-scales could be viewed as a kind of informational closure and coarse-graining (Hoel et al., 2016; Chang et al., 2019). Further, for self-evidencing generative models (Hohwy, 2016; Yufik and Friston, 2016; Kirchhoff et al., 2018), integrating processes underlying SOHM formation would calculate marginal joint posteriors based on specific (Bayesian) beliefs entailed by particular patterns of effective connectivity within and between various synchronous complexes.

[Note: More details on potential mechanisms for SOHM formation and functional consequences can be found in IWMT Revisited (Safron, 2019a) *in the sections: "Phenomenal binding via ESMs (Embodied Self-Models)" and "Mechanisms for integration and workspace dynamics."]*

SOHMs as Dynamic Cores of Integrated Information and Workspaces

With respect to conscious perception, the resonant signal amplification by which SOHMs emerge could potentially contribute to the calculation of highly precise—albeit not necessarily accurate (Hohwy, 2012; Vul et al., 2014)—joint distributions (or maximal a posteriori (MAP) estimates derived thereof). The ability of synchronous complexes to select phase-aligned patterns has clear correspondences with theories of consciousness emphasizing re-entrant signaling (Singer, 2001; Varela et al., 2001; Crick and Koch, 2003; Edelman et al., 2011; Shanahan, 2012; Dehaene, 2014; Grossberg, 2017) and in terms of Bayesian model selection (Hohwy, 2012, 2013), could be understood as promoting winner-take-all dynamics among competing and cooperating inferential flows. SOHMs may also help provide mechanistic bases for "ignition" events accompanying phase transitions in which perception becomes conscious (Dehaene and Changeux, 2011; Friston et al., 2012a; Arese Lucini et al., 2019). IWMT specifically proposes that conscious ignition corresponds to surpassing critical thresholds for SOHM formation via self-synchronized neural activity, thus forming meta-stable complexes as dynamic cores of integrated information, functioning as neuronal global workspaces.

TABLE 3 | Neural frequency bands, their potential roles in predictive processing, and possible experiential consequences.

Frequency band	Role in predictive processing	Potential experiential consequences
Gamma (∼30–120 Hz)	Ascending prediction-errors	Sensory sensitivity and detail
Beta (∼13–30 Hz)	Descending predictions	Perceptual vividness
Alpha (∼8–12 Hz)	Predictions integrated into coherent (egocentric) spatial, temporal, and causal reference frames	Basic phenomenal consciousness
Theta (∼3–7 Hz)	Predictions integrated with internally-generated actions and comparisons among recent (and counterfactual) experiences	Access consciousness, agency, and shaping of phenomenal consciousness via actions
Delta (∼0.5–2 Hz)	Higher-level predictions for active inference unfolding at slower and more inclusive temporal and spatial scales	Unclear; possibly autonoetic consciousness and complex cognition; emotions and feelings, broadly construed as global alterations of states of consciousness and means of aligning spatiotemporal dynamics between mind and world (Northoff and Huang, 2017)

The ability of SOHMs to select aligned patterns may help explain how seemingly definite experiences could emerge from probabilistic world models (Wiese, 2017; Block, 2018; Clark, 2018; Gross, 2018), as opposed to generating a "Bayesian blur," or superposition of possibilities. This hypothesis is consistent with Clark's (2018) suggestion that coherent and precise inference stems from requirements for engaging with environments via sensorimotor couplings (Clark, 2016). Along these lines, by enabling the generation of inferences with rapidity and reliability, SOHMs could afford approximate models capable of guiding action–perception cycles and decision-making (von Uexküll, 1957; Fuster, 2009; Madl et al., 2011; Vul et al., 2014; Linson et al., 2018; Parr and Friston, 2018b). Further, these sensorimotor engagements may promote SOHM formation by providing coherent sources of correlated information, thus affording the possibility of learning even more sophisticated models (Pfeifer and Bongard, 2006; Safron, 2019a,c). IWMT proposes that this continual shaping of behavior based on rich causal world models may be both a major adaptive function of consciousness and a precondition for developing coherent conscious experience. [Note: If consciousness requires semiotic closure Chang et al., 2019 via action–perception cycles, then this cybernetic grounding suggests that systems like plants and insect colonies are unlikely to be conscious, even if capable of sophisticated (but limited) levels of intelligence.]

SOHM dynamics may help to explain many kinds of rhythmic phenomena, such as the fact that oscillations tend to occur at faster rates in organisms with smaller brains (Buzsáki and Watson, 2012); all else being equal, smaller systems are likely to arrive at synchronous equilibria more quickly, with larger systems requiring relatively more time for synchronizing their micro-dynamics. SOHMs may also help to explain why different rhythms (**Table 3**) would be associated with different processes in hierarchical predictive processing (HPP) (Bastos et al., 2015; Sedley et al., 2016; Chao et al., 2018), where faster gamma oscillations communicate bottom-up prediction-errors 'calculated' by local microcircuits, and where slower beta oscillations generate top-down predictions via integrating information (i.e., accumulating model evidence) from more spatially-extended sources. These beta complexes may potentially be organized via nesting within even larger and slower-forming SOHMs, such as those generated at alpha, theta, and delta frequencies. This cross-frequency phase coupling (Canolty and Knight, 2010) could allow for the stabilization of multi-scale dynamics within HPP, with increasing levels of hierarchical depth affording modeling of complex and temporally extended causes (Friston et al., 2017c). Hierarchical nesting of SOHMs could allow modeling to simultaneously (and synergistically) occur at multiple levels of granularity, thus affording both global stability (Humphrey, 2017) and fine-grained adaptive control as overall systems couple with their environments.

If SOHMs integrate information in the ways suggested here—marginalizing over synchronized components—then the largest SOHM of a system would generate a joint posterior (or estimate derived thereof) over all smaller SOHMs contained within its scope. These encompassing SOHMs would integrate information across heterogeneous processes, as well as affording unified sources of control energy for system evolution. These maximal SOHMs could generate estimates of overall organismic states, thus forming dynamic cores of integration for perception and action, potentially enabling autonomous control by integrated self-processes. Further, privileged positions of maximal SOHMs with respect to network centrality (Aadithya et al., 2010) and modeling capacity could promote directional entrainment of smaller complexes, thus promoting coherent agentic action selection.

For biological systems, the dynamics within maximal SOHMs may have the clearest correspondences with events unfolding at organismic scales. For organisms such as *C. elegans*, these dynamics might unfold at the frequencies of locomotory eigenmodes, potentially concentrated in a core of richly connected nodes (Towlson et al., 2013), thus allowing enslavement of a worm's peripheral pattern generators by predictive models coordinating the enaction of coherent movement vectors. For organisms such as *Homo sapiens*, these dynamics might unfold at the frequencies of real and imagined sensorimotor contingencies (Elton, 2000; O'Regan and Noë, 2001; Tani, 2016; Chen et al., 2017; Prinz, 2017; Zadbood et al., 2017; Baldassano et al., 2018; Chang et al., 2019), potentially concentrated along deep portions of cortical generative models, thus allowing enslavement of an individual's sensorium and effectors by rich causal models of self and world. Whether

in worms or humans, SOHMs would entail joint posteriors (or associated maximal estimates) from probabilistic models for embodied agents and the environments with which they couple. In these ways, Maximal SOHMs may be coextensive with both maxima of integrated information (i.e., MICS) and global workspaces. However, while SOHMs with the greatest amount of irreducible integrated information may correspond to basic phenomenal consciousness (e.g., complexes centered on posterior cortices), organization into an even larger (albeit possibly less irreducibly integrated) synchronous complex involving the frontal lobes may be required for access consciousness and agentic control.

A multi-level understanding of SOHMs in terms of neuronal dynamics and probabilistic inference suggests that we should expect these complexes to form over subnetworks with coherent mutual information, which is more likely if patterns of effective connectivity entail coherent and well-evidenced world models. With respect to loopy message passing for approximate inference (Koller and Friedman, 2009; Friston et al., 2017b), these coherent models may have a (circular) causal significance in that they would be more likely to provide consistent inferential flows, and so be more likely to first converge upon stable posteriors, and so be more likely to dominate rounds of Bayesian model selection. Notably, this kind of convergence is more likely for Bayesian networks that balance integration and differentiation—associated with high phi (Marshall et al., 2016)—and this is precisely what is observed for "rich club" connectivity cores (Sporns, 2013; Mišić et al., 2015; Cohen and D'Esposito, 2016; Mohr et al., 2016). Further, high degrees of re-entrant connectivity and potential for recurrent dynamics suggests that these richly connected networks are particularly likely to serve as loci of "ignition" events in global workspace models (Dehaene and Changeux, 2011; Shanahan, 2012). Finally, considering that integrated information reflects a system's ability to exert cause–effect power over itself, SOHMs may be particularly likely to form along high phi networks.

IWMT and Maximizing SOHMs: Bringing Forth Worlds of Experience

A maximal SOHM—as a MICS and MICE repertoires—within a brain may center on posterior cortices, and in particular the temporoparietal junction (Graziano, 2019) and posteromedial cortices (PMCs) (O'Reilly et al., 2017), with synchronizing complexes forming at alpha frequencies generating basic phenomenal consciousness. Nesting of these alpha rhythms within theta frequencies may further allow for coupling with the frontal lobes and hippocampal complex, thus affording goal-directed and access consciousness from global workspace dynamics. IWMT's focus on PMCs and alpha frequencies (as synchronizing manifolds) is based on both the types of information available to these systems/processes (Papez, 1937; Jann et al., 2009; Gramann et al., 2010; Knyazev et al., 2011; Damasio, 2012), as well as empirical associations with attention and working memory (Palva and Palva, 2011; Kerr et al., 2013; Michalareas et al., 2016; Sato et al., 2018; Bagherzadeh et al., 2019). PMCs receive information from upper levels of each sensory hierarchy, as well as the position of an organism in space, including head-direction information. This information is likely a prerequisite for organizing perception into egocentric reference frames (Brewer et al., 2011, 2013; Guterstam et al., 2015; Li et al., 2018; Smigielski et al., 2019). In line with models in which consciousness depends on projective geometry (Rudrauf et al., 2017; Williford et al., 2018), a stable source of egocentric perspective may represent a practically necessary precondition for there to be "something that it is like:" i.e., the ability to generate models with spatial, temporal, and causal coherence for system and world.

IWMT focuses on space, time (i.e., relative dynamics in space), and cause (i.e., predictable regularities in these dynamics), but wholistic self-processes (Damasio, 2012; Humphrey, 2017) may also be essential for developing world models capable of generating coherent subjectivity. Self-processes may be practically necessary for consciousness because the integration of large-scale brain activity may be required for the coherent regulation of action–perception cycles, and thereby cybernetic sense-making. Self-processes could allow for selection of specific models on the basis of relevance (Shanahan and Baars, 2005; Davey and Harrison, 2018; Linson et al., 2018; Hattori et al., 2019), with stable self-models extending this organization across time (Dennett, 1992; Hirsh et al., 2013; Buonomano, 2017), thereby enabling the learning required to construct experienceable world models. In brief, IWMT proposes that Kant's preconditions for judgment are also necessary preconditions for consciousness (Northoff, 2012; De Kock, 2016). While PMCs may be sufficient for basic phenomenal consciousness, larger complexes may be required for certain kinds of higher-order cognition, including access and autonoetic consciousness (Brown et al., 2019; LeDoux, 2019; Shea and Frith, 2019). This integration of action with perception is likely crucial for agentic planning and the counterfactual simulations upon which it is based (Hassabis and Maguire, 2009; MacKay, 2019), without which the development of coherent world models may be impossible (De Kock, 2016; Friston, 2017).

To summarize (**Table 4**), in systems where synchrony both emerges from and facilitates coherent message passing, SOHMs enable both workspace dynamics and high degrees of meaningful informational integration, where meaning is a difference that makes a difference to the ability of a system to survive and achieve its goals. However, integrated information and workspaces only entail consciousness when applied to systems that can also be understood as Bayesian belief networks, where beliefs have coherence because they have actual semantic content by virtue of evolving through interactions with a coherently structured (and so semi-predictable) world. Without those meaningful external connections, systems could have arbitrarily large amounts of integrative potential, but there still may be nothing that it is like to be such systems.

*[Note: For some testable hypotheses related to these ideas, please refer to **Supplementary Material**.]*

TABLE 4 | Integrating IIT with the FEP-AI framework and IWMT's model of communication through coherence via SOHM dynamics.

Integrated Information Theory (IIT) axioms and postulates	Integration with the Free Energy Principle and Active Inference (FEP-AI) Framework	Integration via Self-Organizing Harmonic Modes (SOHMs): Eigenmodes of effective connectivity and synchronization manifolds
Intrinsic existence:		
Systems exert C–E power on themselves and the sub-systems of which they are composed. Sub-systems exert C–E power on themselves and the larger systems of which they are a part. C–E power exists at particular spatial and temporal grains.	Systems are describable as PGMs, where graphs express conditional dependence structure between sub-components. All systems and sub-systems possess defining MBs, the boundaries of which establish conditional independence between internal and external states. MB internal states can only interact with themselves, or with external states via MBs. Persisting systems preserve their MBs by exerting C–E power both on themselves and other systems.	SOHMs (and their MB boundaries) form as systems and sub-systems interact with both themselves and other systems at particular spatial and temporal grains. SOHMs influence how systems as wholes are likely to interact with both themselves and other systems at varying levels of granularity. SOHMs are both consequences and causes of the processes that generate them, both emerging from and determining the C–E power that systems exert on themselves and other systems.
Composition:		
Systems are composed of sub-systems with particular inter-relations. Structured inter-relations determine the specific C–E power of systems on sub-systems, which exert C–E power on each other.	PGMs are composed of connected elements with particular components differentially contributing to joint probability distributions. Graph structures define relations of conditional dependence and independence, so determining inferential flows within and between MBs (i.e., marginalization and message passing). Persisting MB compositions are generative models for those particular compositions.	Particular system compositions influence the dynamics of SOHM formation, which, in turn, influence patterns of effective connectivity between and within system sub-components. Subnetworks along which SOHMs form determine how C–E power flows on different timescales, including with respect to SOHM formation processes. SOHMs have specific spatial and temporal extents, so defining systems and sub-systems in terms of particular inter-relations.
Information:		
Systems have specific compositions that are differentiated from other possible compositions. C–E repertoire: probability distribution over all permutations of possible causes and effects that a system could exert on itself.	MB-defined dependency relations specify inferential properties of PGMs, including probability distributions and estimates for likely causes of present observations, given past observations. Mappings from observations to likely causes define systems as generative models.	Specific combinations of SOHMs and their particular compositions influence (and are influenced by) effective connectivity within and between systems, so specifying the particular information content of those systems. By promoting communication through coherence, MB-bounded SOHMs can implement marginalization over sub-networks and organize message passing and/or belief propagation.
Conceptual structure:		
Mapping of C–E repertoires onto an abstract metric space, specifying particular causal properties.	Persisting systems generate themselves as particular densities, so providing mutual information between past and future states, and between internal and external states of MB-bound systems.	Different systems will have different SOHMs, so generating inferences that are differentiated from other systems in which different groups of elements would be included within synchronizing manifolds.
Integration:		
Systems are unified in terms of being irreducible to independent subcomponents. This irreducibility can be quantified (phi) by comparing C–E repertoires before and after systems are divided by a minimally disruptive partitioning, known as a "minimal information partition" (MIP).	All components of MB-bounded sub-graphs from PGMs (differentially) contribute to integrating—literally, calculating integrals for—associated marginal joint probability distributions. Persisting systems are unified (to varying degrees); all components contribute to self-evidencing (to varying extents). By quantifying the integrated complexity of system-internal C–E power, the phi of an MB-bound set will correlate with the marginal likelihood (or negative free energy) associated with particular self-evidencing systems.	SOHMs are unified (to varying degrees); all components of self-interacting systems contribute (to varying extents) to the emergence of its particular eigenmodes. If SOHMs influence and are influenced by the particular configuration of a system, then any alteration will result in different patterns of effective connectivity. If SOHMs promote information transmission, then any SOHM modification will change inferences, where the least of these alterations would constitute a MIP.
Exclusion:		
Systems have definite boundaries with respect to their ability to exert C–E power over particular spatial and temporal grains. IIT identifies intrinsically existing systems as complexes, specifying maximally irreducible conceptual structures (MICS) and associated maximally irreducible cause-effect (MICE) repertoires.	PGMs represent multiple possibilities, but they can also generate precise posterior distributions and discrete estimates of likely parameter values. Larger systems can integrate marginal probabilities from MB-bounded sub-systems, so integrating more information into models. If phi promotes self-generation, then boundaries for maximal complexes would correspond to boundaries for (free-energy-minimizing) systems generating maximal self-model evidence, with maximal potential influences on overall system evolution.	The specific temporal and spatial scales governing SOHM formation will constrain opportunities for influencing the evolution of these self-synchronizing attracting manifolds. The MB boundaries of SOHMs will define which dynamics are capable of contributing to joint inference to which degrees. Theoretically, rapidly forming and strongly synchronizing SOHMs could entail precise joint probabilities, or maximum a posteriori (MAP) estimates derived thereof.

C-E, Cause-effect; PGM, Probabilistic graphical model; MB, Markov blanket.

DISCUSSION: TOWARD SOLVING THE ENDURING PROBLEMS OF CONSCIOUSNESS (AND AI?)

[Note: More details on computational principles and systems likely to be associated with consciousness can be found in IWMT Revisited (Safron, 2019a) *in the sections, "Machine learning architectures and predictive processing models of brain and mind" and "Consciousness: Here, There, but Not Everywhere."]*

Autoencoders, Predictive Processing, and the Conscious Turbo Code

Helmholtz (1878) is often viewed as providing the first clear description of perception as inference:

> "*Objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism.*"

Dayan, Hinton, Neal, and Zemel (Dayan et al., 1995) constructed machine learning systems based on these principles, trained using cost functions based on Helmholtz free energy. These kinds of architectures can be trained to handle noisy inputs or infer missing data, with more recent versions being able to generate completely novel combinations of features. These are all aspects of conscious (and unconscious) perception and have many commonalities with HPP within FEP-AI.

Variational autoencoders (Kingma and Welling, 2014) are composed of encoders and generative decoders connected by low-dimensional bottlenecks, where encoders learn to compress input data into reduced-dimensionality feature spaces, and where decoders learn to use these latent features to infer likely details of higher-dimensional data. HPP models of sensory cortices (**Figure 2**; **Table 5**) may be approximated as disentangled variational autoencoders, where encoders and decoders are constituted by respective hierarchies of superficial and deep pyramidal neurons (Kanai et al., 2019). However, rather than training solely based on divergences between respective input and output layers of encoder and decoder networks, prediction-error is minimized at all levels simultaneously based on comparisons between time-varying sensory observations and internally-generated predictions. HPP in brains further involves multiple interacting autoencoding hierarchies, with connections being particularly strong in deeper association cortices—corresponding to reduced dimensionality latent spaces—thus affording synergistic inferential power with shared priors from multi-modal sensory integration and world modeling.

IWMT proposes that connections between the low-dimensionality bottlenecks from various modalities may form an auto-associative network supporting loopy belief propagation—or message passing—thus constituting a turbo code (Berrou and Glavieux, 1996), and hence approaching the Shannon limit with respect to optimality in communicating information over noisy channels (**Figure 3**; **Table 6**). [Note: While any instantiation of loopy belief propagation may be understood as realizing a turbo code, IWMT specifically suggests that a broad network of cross-modal effective connectivity is required for coherent integrated world modeling.] This framing of HPP in terms of autoencoders and turbo codes could provide a computational analog for neural systems underlying consciousness: a reduced-dimensionality representational bottleneck that extracts the most important details from sensory data, and which affords inferential synergy by providing a workspace where specialist models can be combined, integrated, and then rebroadcast. [Note: HPP dimensionality-reduction may have relevance to the sketch-like nature of awareness proposed in Graziano's Attention Schema Theory (Graziano, 2013, 2019).] According to IWMT, coherent self-world modeling likely also requires organizing this information into spatiotemporal trajectories, as afforded by the hippocampal system and machine learning architectures that attempt to reproduce its functioning (Fraccaro et al., 2017; Ha and Schmidhuber, 2018; Whittington et al., 2018; Wu et al., 2018), and as suggested by impaired counterfactual modeling with medial temporal lobe damage (Hassabis and Maguire, 2009; MacKay, 2019).

As Bengio (2017) has suggested with his work on the "consciousness prior," the reduced dimensionality of these (disentangled) features may be well-suited for identifying major axes of meaningful variations in the world, such as those involved in the kinds of causal processes we can manipulate and perceive, and which can also be mapped onto linguistic systems. This later affordance has relevance to Higher-Order Theories of consciousness, including those emphasizing agentic modeling and social communication (Metzinger, 2010; Graziano, 2013; Rudrauf et al., 2017; Brown et al., 2019; Shea and Frith, 2019).

The thalamocortical system enabling dynamic cores of integration and conscious workspaces first evolved hundreds of millions of years before these higher-order processes existed (Edelman, 2004). These richly connected subnetworks enable high-bandwidth message passing—as likely required for realizing turbo codes in biological systems—but are also metabolically expensive, consuming nearly 50% of cortical metabolism in humans (Heuvel et al., 2012). However, part of the way these energetic costs may be justified is by (a) reducing the number of (noisy) neuronal signal transactions required to achieve adequately reliable perceptual inference, (b) enhancing the speed of model selection for the sake of fine-grained control, and (c) allowing for imagination-based planning and causal reasoning (Pearl and Mackenzie, 2018). Rich-club connected subnetworks can even be found in *C. elegans* with their 302 neurons (Towlson et al., 2013). This could be taken to imply that consciousness is nearly a billion years old, but IWMT suggests that this is likely a mistaken inference, as deep hierarchies may be required for generating coherent experience.

Conscious AI?

IWMT does not suggest that consciousness corresponds to either the output layers of generative models as currently used in machine learning or the processes calculating those outputs. Although architectures with self-attention mechanisms have been implemented with great success (Kovaleva et al., 2019), the outputs of such systems tend to be functionally disconnected from each other, as well as the processes by which they are

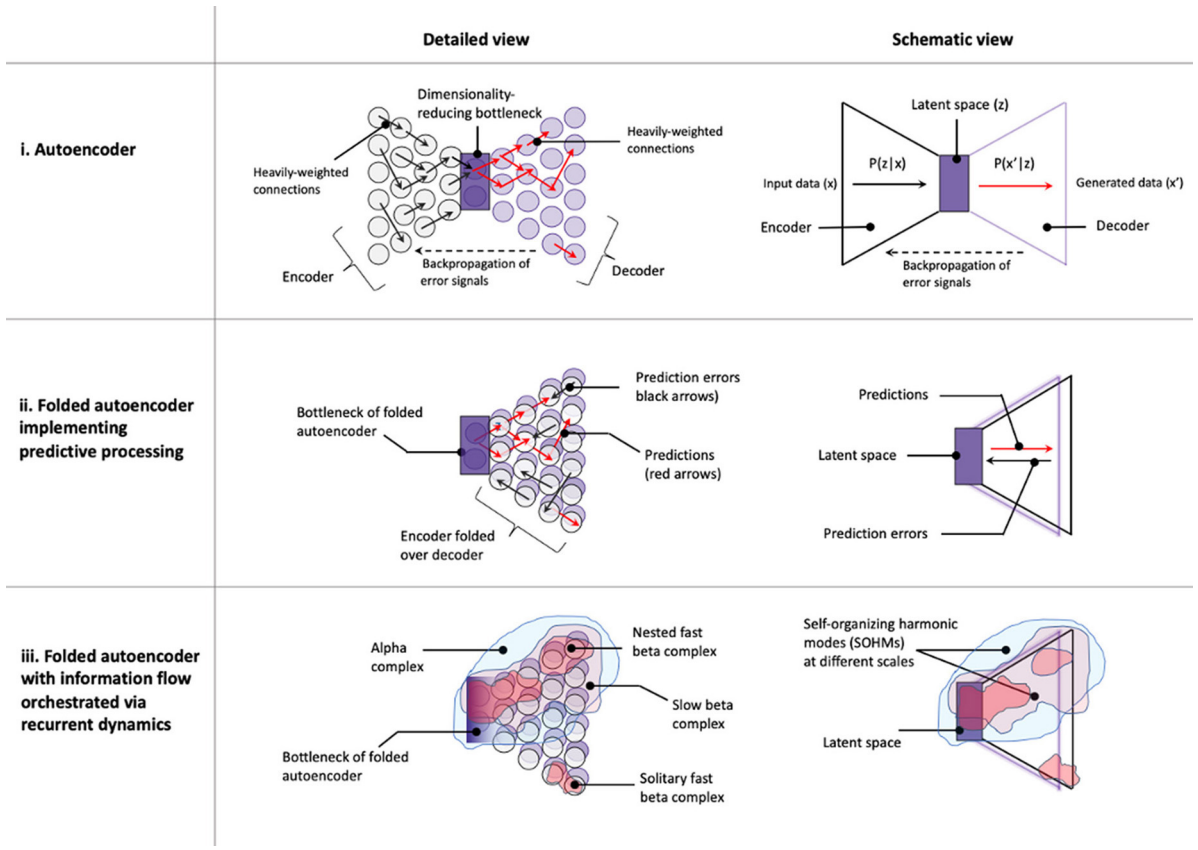


FIGURE 2 | Sparse folded variational autoencoders with recurrent dynamics via self-organizing harmonic modes (SOHMs).

(i) Autoencoder.
An autoencoder is a type of artificial neural network that learns efficient representations of data, potentially including a capacity for generating more complete data from less complete sources. The encoder compresses input data over stages of hierarchical feature extraction, passes it through a dimensionality-reducing bottleneck and into a decoder. The decoder attempts to generate a representation of the input data from these reduced-dimensionality latent representations. Through backpropagation of error signals, connections contributing to a more inaccurate representation are less heavily weighted. With training, the decoder learns how to generate increasingly high-fidelity data by utilizing the compressed (and potentially interpretable) feature representations encoded in the latent space of the bottleneck portion of the network. In the more detailed view on the left, black arrows on the encoder side represent connections contributing to relatively high marginal likelihoods for particular latent feature space representations, given connection weights and data. Red arrows on the decoder side represent connections with relatively high marginal likelihoods for those reconstructed features, given connection weights and latent space feature hypotheses. While these autoencoders are fully connected dense networks, particular connections are depicted (and associated probabilities discussed) because of their relevance for predictive processing. Note: Although the language of probability theory is being used here to connect with neurobiologically-inspired implementations, this probabilistic interpretation—and links to brain functioning—is more commonly associated with variational autoencoders, which divide latent spaces into mean and variance distributions parameterized by stochastic sampling operations in generating likely patterns of data, given experience.

(ii) Folded autoencoder implementing predictive processing.
In this implementation of predictive processing, autoencoders are 'folded' at their low-dimensionality bottlenecks—such that corresponding encoding and decoding layers are aligned—with decoding hierarchies (purple circles) depicted as positioned underneath encoding hierarchies (gray circles). Within a brain, these decoding and encoding hierarchies may correspond to respective populations of deep and superficial pyramidal neurons (Bastos et al., 2012). In the figure, individual nodes represent either units in an artificial network—or groups of units; e.g., capsule networks (Kosiorek et al., 2019)—or neurons (or neuronal groups; e.g., cortical minicolumns) in a brain. Predictions (red arrows) suppress input signals when successfully predicted, and are depicted as traveling downwards from representational bottlenecks (corresponding to latent spaces) along which autoencoding networks are folded. Prediction errors, or observations for a given level (black arrows) continue to travel upwards through encoders unless they are successfully predicted, and so "explained away." Data observations (i.e., prediction errors) are depicted as being sparser relative to high-weight connections in the (non-folded) encoding network presented above, where sparsity is induced via predictive suppression of ascending signals. This information flow may also be viewed as Bayesian belief propagation or (marginal) message passing (Friston et al., 2017b; Parr et al., 2019). In contrast to variational autoencoders in which training proceeds via backpropagation with separable forward and backward passes—where cost functions both minimize reconstruction loss and deviations between posterior latent distributions and priors (usually taking the form of a unit Gaussian)—training is suggested to occur (largely) continuously in predictive processing (via folded autoencoders), similarly to recent proposals of target propagation (Hinton, 2017; Lillicrap et al., 2020). Note: Folded autoencoders could potentially be elaborated to include attention mechanisms, wherein higher-level nodes may increase the information gain on ascending prediction-errors, corresponding to precision-weighting (i.e., inverse variance over implicit Bayesian beliefs) over selected feature representations.

(Continued)

FIGURE 2 | (iii) Folded autoencoder with information flows orchestrated via recurrent dynamics.
This row shows a folded autoencoder model of a cortical hierarchy, wherein neuronal oscillations mediate predictions—potentially orchestrated by deep pyramidal neurons and thalamic (and striatal) relays—here characterized as self-organizing harmonic modes (SOHMs). This paper introduces SOHMs as mechanisms realizingsynchronization manifolds for coupling neural systems (Palacios et al., 2019), and sources of coherent neuronal oscillations and evidence accumulation for predictive processing. Depending on the level of granularity being considered, these predictive oscillations could either be viewed as traveling or standing waves (i.e., harmonics). SOHM-based predictions are shown as beta oscillations forming multiple spatial and temporal scales. These predictive waves may be particularly likely to originate from hierarchically higher levels—corresponding to latent spaces of representational bottlenecks—potentially due to a relatively greater amount of internal reciprocal connectivity, consistent information due to information aggregation, or both. SOHMs may also occur at hierarchically lower levels due to a critical mass of model evidence accumulation allowing for the generation of coherent local predictions, or potentially on account of semi-stochastic synchronization. Faster and smaller beta complexes are depicted as nested within a larger and slower beta complex, all of which are nested within a relatively larger and slower alpha complex. Note: In contrast to standard machine learning implementations, for this proposal of predictive processing via folded autoencoders (and SOHMs), latent space is depicted as having unclear boundaries due to its realization via recurrent dynamics. Further, inverse relationships between the spatial extent and speed of formation for SOHMs are suggested due to the relative difficulties of converging on synchronous dynamics within systems of various sizes; theoretically, this mechanism could allow for hierarchical modeling of events in the world for which smaller dynamics would be expected to change more quickly, and where larger dynamics would be expected to change more slowly.

TABLE 5 | Proposed correspondences between features of variational autoencoders and predictive processing.

Variational autoencoder features	Proposed correspondences in predictive processing
Encoder network	Ascending hierarchy of superficial pyramidal neurons; Message-passing at gamma frequencies
Generative decoder network	Descending hierarchy of deep pyramidal neurons; Beliefs propagated at beta frequencies
Reduced dimensionality bottleneck	Association cortices and deeper portions of generative models; Estimates calculated at beta, alpha, and theta frequencies
Mean vectors	Activity levels for neuronal populations at different parts of hierarchy
Variance vectors	Neuronal population activity variability
Sampling from latent feature space	Large-scale synchronous complexes at beta, alpha, and theta frequencies; "ignition" events
Training: minimizing reconstruction loss between input layer of encoder and output layer of generative decoder; also minimizing divergence from unit Gaussian, parameterized by disentangling parameter	Training: minimizing precision-weighted prediction-errors at all layers simultaneously; precision-weighting as analogous to disentanglement hyperparameter; many mechanisms including synchronous gain control and diffuse neuromodulatory systems
Potential for sequential organization via recurrent network controllers (Ha and Schmidhuber, 2018)	Organization of state transitions by hippocampal system and frontal cortices (Koster et al., 2018)

generated. This is not the case for brains, for which IWMT proposes that joint posteriors and estimates (and samples derived thereof) are calculated via spreading neuronal activity, where message-passing/belief-propagation is promoted (or scheduled) via synchronous dynamics (i.e., SOHMs). As opposed to current generations of generative models, the functioning of these synchronized subnetworks (and the calculations they entail) span multiple levels of hierarchical depth, with bidirectional linkages to generative processes involving models with spatial, temporal, and causal coherence for system and world.

Further, the anatomical and functional directedness of neuronal connections at any point in time contain information that will bias future dynamics, so influencing likelihoods with which meta-stable regimes are subsequently produced. If these networks are altered according to principles of spike-timing dependent plasticity, and if systems develop in the context of embodied agents interacting with their environments, then these state transitions are likely to contain coherent information reflecting causal world structures (Hayek, 1952; Markram et al., 2011; Lakoff, 2014). In these ways and more (e.g. recurrent dynamics persisting across SOHM-formation events), each quale state would functionally connect and constrain future quale states based on past quale states. Further, biological neural networks do not generate feature maps as isolated vectors over stimulus dimensions, but as vectors coupled over multiple levels of hierarchical depth, via neuronal dynamics. Thus, consciousness may be entailed by the functioning of a probabilistic model that generates tensors in neuronal (and representational) phase space, specifying joint posteriors (or estimates derived thereof), where that which is being modeled/estimated is the causes of sensation. If this is the type of mathematical object that corresponds to subjective experience, then substantial progress may have been made toward solving the Hard problem of consciousness.

Conclusion: Toward (Dis-) Solving the Meta-Problem by Solving the Hard Problem

How could there be "something that it is like" to be a physical system or entailed mathematical object? IWMT suggests that this question may be satisfyingly answered if such a system can calculate—or probabilistically infer—sequences of sensorimotor states. Perhaps intuitively, such a sequential unfolding would have more of a resemblance to the flowing of the stream of consciousness for the kinds of embodied–embedded agents that we are. If models can generate particular combinations of information present within and between sensory modalities, then

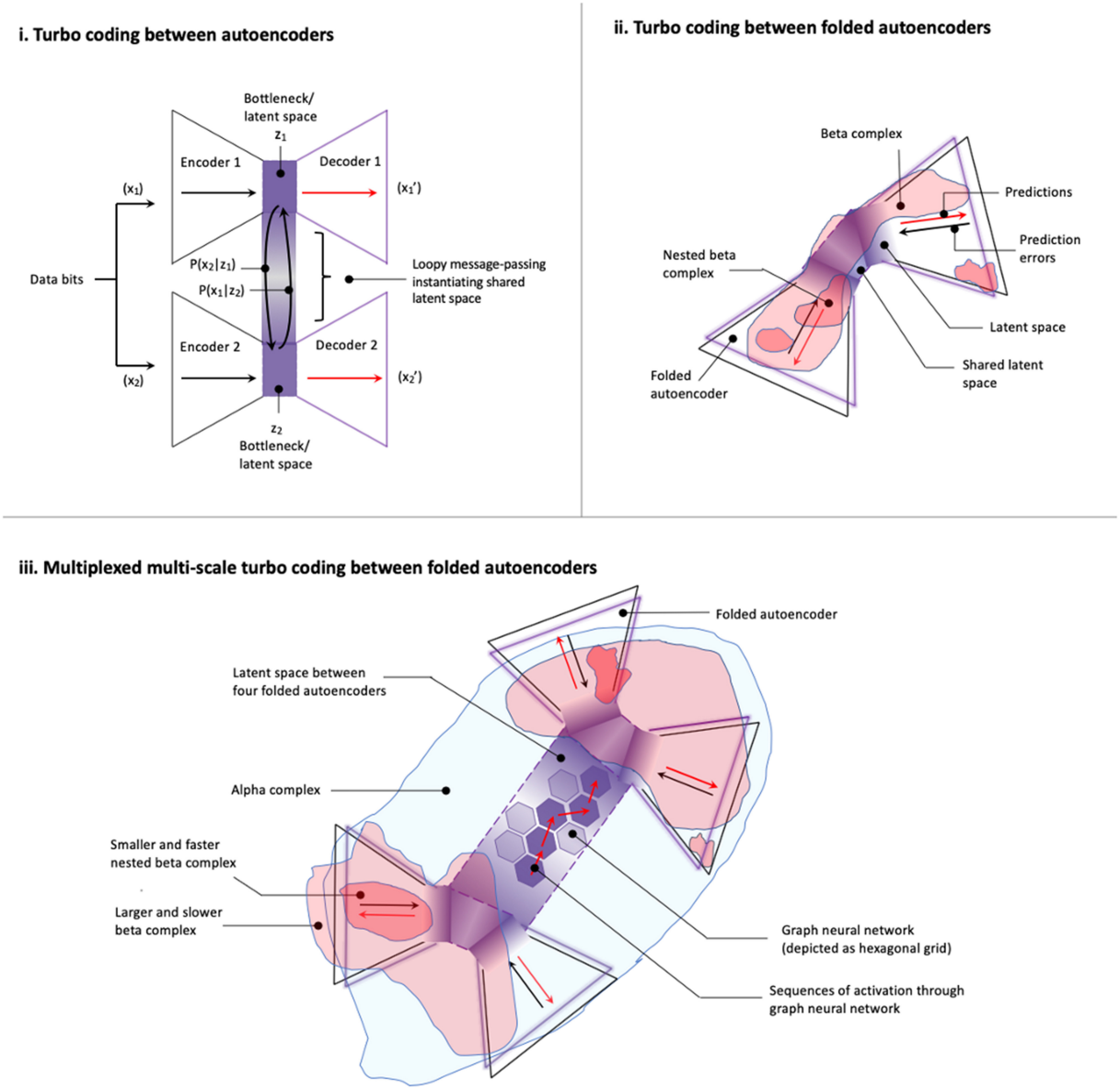


FIGURE 3 | Cortical turbo codes.

(i) Turbo coding between autoencoders.
Turbo coding allows signals to be transmitted over noisy channels with high fidelity, approaching the theoretical optimum of the Shannon limit. Data bits are distributed across two encoders, which compress signals as they are passed through a dimensionality reducing bottleneck—constituting a noisy channel—and are then passed through decoders to be reconstructed. To represent the original data source from compressed signals, bottlenecks communicate information about their respective (noisy) bits via loopy message passing. Bottleneck z_1 calculates a posterior over its input data, which is now passed to Bottleneck z_2 as a prior for inferring a likely reconstruction (or posterior) over its data. This posterior is then passed back in the other direction (Bottleneck z_2 to Bottleneck z_1) as a new prior over its input data, which will then be used to infer a new posterior distribution. This iterative Bayesian updating repeats multiple times until bottlenecks converge on stable joint posteriors over their respective (now less noisy) bits. IWMT proposes that this operation corresponds to the formation of synchronous complexes as self-organizing harmonic modes (SOHMs), entailing marginalization over synchronized subnetworks—and/or precision-weighting of effectively connected representations—with some SOHM-formation events corresponding to conscious "ignition" as described in Global Neuronal Workspace Theory (Dehaene, 2014). However, this process is proposed to provide a means of efficiently realizing (discretely updated) multi-modal sensory integration, regardless of whether "global availability" is involved. Theoretically, this setup could allow for greater data efficiency with respect to achieving inferential synergy and minimizing reconstruction loss during training in both biological and artificial systems. In terms of concepts from variational autoencoders, this loopy message passing over bottlenecks is proposed to entail discrete updating and maximal a posteriori (MAP) estimates, which are used to parameterize semi-stochastic sampling operations by decoders, so enabling the iterative generation of likely patterns of data, given past experience (i.e., training) and present context (i.e., recent data preceding turbo coding). Note: In turbo coding as used in industrial applications such as enhanced telecommunications, loopy message passing usually proceeds between interlaced decoder networks; within cortex, turbo coding could potentially occur with multiple (potentially nested) intermediate stages in deep cortical hierarchies.

(Continued)

FIGURE 3 | (ii) Turbo coding between folded autoencoders.
This panel shows turbo coding between two folded autoencoders connected by a shared latent space. Each folded autoencoder sends predictions downwards from its bottleneck (entailing reduced-dimensionality latent spaces), and sends prediction errors upwards from its inputs. These coupled folded autoencoders constitute a turbo code by engaging in loopy message passing, which when realized via coupled representational bottlenecks is depicted as instantiating a shared latent space via high-bandwidth effective connectivity. Latent spaces are depicted as having unclear boundaries—indicated by shaded gradients—due to their semi-stochastic realization via the recurrent dynamics. A synchronous beta complex is depicted as centered on the bottleneck latent space—along which encoding and decoding networks are folded—and spreading into autoencoding hierarchies. In neural systems, this spreading belief propagation (or message-passing) may take the form of traveling waves of predictions, which are here understood as self-organizing harmonic modes (SOHMs) when coarse-grained as standing waves and synchronization manifolds for coupling neural systems. Relatively smaller and faster beta complexes are depicted as nested within—and potentially cross-frequency phase coupled by—this larger and slower beta complex. This kind of nesting may potentially afford multi-scale representational hierarchies of varying degrees of spatial and temporal granularity for modeling multi-scale world dynamics. An isolated (small and fast) beta complex is depicted as emerging outside of the larger (and slower) beta complex originating from hierarchically higher subnetworks (hosting shared latent space). All SOHMs may be understood as instances of turbo coding, parameterizing generative hierarchies via marginal maximum a posteriori (MAP) estimates from the subnetworks within their scope. However, unless these smaller SOHMs are functionally nested within larger SOHMs, they will be limited in their ability to both inform and be informed by larger zones of integration (as probabilistic inference).

(iii) Multiplexed multi-scale turbo coding between folded autoencoders.
This panel shows turbo coding between four folded autoencoders. These folded autoencoders are depicted as engaging in turbo coding via loopy message passing, instantiated by self-organizing harmonic modes (SOHMs) (as beta complexes, in pink), so forming shared latent spaces. Turbo coding is further depicted as taking place between all four folded autoencoders (via an alpha complex, in blue), so instantiating further (hierarchical) turbo coding and thereby a larger shared latent space, so enabling predictive modeling of causes that achieve coherence via larger (and more slowly forming) modes of informational integration. This shared latent space is illustrated as containing an embedded graph neural network (GNN) (Liu et al., 2019; Steppa and Holch, 2019), depicted as a hexagonal grid, as a means of integrating information via structured representations, where resulting predictions can then be propagated downward to individual folded autoencoders. Variable shading within the hexagonal grid-space of the GNN is meant to indicate degrees of recurrent activity—potentially implementing further turbo coding—and red arrows over this grid are meant to indicate sequences of activation, and potentially representations of trajectories through feature spaces. These graph-grid structured representational spaces may also afford reference frames at various levels of abstraction; e.g., space proper, degrees of locality with respect to semantic distance, abductive connections between symbols, causal relations, etc. If these (alpha- and beta-synchronized) structured representational dynamics and associated predictions afford world models with spatial, temporal, and causal coherence, these processes may entail phenomenal consciousness. Even larger integrative SOHMs may tend to center on long-distance white matter bundles establishing a core subnetwork of neuronal hubs with rich-club connectivity (van den Heuvel and Sporns, 2011). If hippocampal-parietal synchronization is established (typically at theta frequencies), then bidirectional pointers between neocortex and the entorhinal system may allow decoders to generate likely patterns of data according to trajectories of the overall system through space and time, potentially enabling episodic memory and imagination. If frontal-parietal synchronization is established (potentially involving theta-, alpha-, and beta- synchrony), these larger SOHMs may also correspond to "ignition" events as normally understood in Global Neuronal Workspace Theory, potentially entailing access consciousness and volitional control.

TABLE 6 | Proposed correspondences between turbo coding in artificial neural networks and biological neural dynamics.

Turbo codes in artificial neural networks	Proposed correspondences in brains
Take data and distribute bits over two encoder–decoder networks.	Each sensory modality can be modeled as a noisy channel.
Generate a posterior probability estimate of the signal in one of the networks.	Within modalities, bottom-up updated states of deeper hierarchical levels calculate local posteriors (possibly taking the form of locally synchronized fast beta complexes).
Take the posterior from this network and propagate that belief as a prior to inform the calculation of a joint posterior for the other network.	Between modalities, auto-associative linkages across deeper hierarchical levels allow posteriors to be shared as empirical priors (possibly taking the form of larger and slower beta complexes).
Pass this message back to the original network as priors to inform the calculation of a new posterior.	Modalities are likely to be reciprocally connected, particularly in proximity to association cortices.
Repeat steps 3 and 4 until loopy belief propagation converges.	The formation of cross-modal synchronized complexes (at slower beta, alpha, and theta) frequencies may entail loopy message passing across modalities via self-organizing harmonic modes (SOHMs).
Result: Highly reliable data transmission even under highly noisy circumstances.	Result: Highly reliable perceptual inferences from noisy and ambiguous sensory information.

we may finally begin to have prima facie reasons to believe that such processes could generate subjective experience.

Global workspaces have been analogized as functioning as (non-Cartesian) theaters (Dehaene, 2014) in which information is rendered visible to otherwise isolated modules, with attention acting as a "spotlight" prioritizing some contents over others. Similar metaphors for awareness have been used by Crick and Koch (2003) with their neuronal coalitions model and also by Hobson and Friston (2016) in suggesting that frontal lobe ensembles produce awareness when they "look" at posterior sensory information. While the implication of some sort of little person in the brain, or homunculus, is nearly universally reviled, this dismissal may be a significant part of the Hard problem's intractability. That is, in attempting to do away with homunculi, cognitive science may have lost track of the importance of both embodiment and centralized control structures. If "cognition" is primarily discussed in the abstract, apart from its embodied–embedded character, then it is only natural that explanatory gaps between brain and mind should seem unbridgeable. IWMT, in contrast, suggests that many quasi-Cartesian intuitions may be partially justified. As discussed in Safron (2019a,c), brains may not only infer mental spaces, but they may further populate these spaces with body-centric representations of sensations and actions at various degrees of

detail and abstraction. From this view, not only are experiences re-presented to inner experiencers, but these experiencers may take the form of a variety of embodied self-models with degrees of agency. In these ways, IWMT situates embodiment at the core of both consciousness and agency, so vindicating many (but not all) folk psychological intuitions.

With respect to the meta-problem, one could imagine postulating a "Hard problem" of generative models in machine learning, for which an unbridgeable explanatory gap may be perceived between the remarkable ability of these architectures to generate rich and novel stimuli (e.g., an "imagined" face), contrasted with the determinism of their underlying computations. Yet this seemingly intractable problem could then be solved via deeper technical understanding. IWMT proposes that this epistemic situation may be analogous to the one we face with consciousness. Rather than the "Hard problem" being reduced to many "easy problems"—and so being (dis-)solved as we discover we were asking the wrong question—it may be the case of this most challenging and profound problem actually being solved through the discovery of sufficiently powerful bridging principles. IWMT suggests such principles may be finally available by using FEP-AI to integrate leading theories of consciousness.

AUTHOR CONTRIBUTIONS

AS conceived and developed this theoretical framework, wrote the entirety of this manuscript, and created all tables and figures therein.

FUNDING

Support for funding the open-access fee was provided by Frontiers and Indiana University.

ACKNOWLEDGMENTS

I would like to thank Karl Friston, Giulio Tononi, Anil Seth, Andy Clark, Selen Atasoy, and Morten Kringelbach for providing me with opportunities to discuss and develop these ideas.

I would also like to extend my thanks to Larissa Albantakis for advising me with respect to IIT. (Note: It should be remembered that IIT attempts to address the Hard problem by starting from axioms of phenomenology and proceeding to formally specify analytic approaches. The functionalist and mechanistic-convergence approach attempted in this manuscript are ultimately incompatible in terms of IIT's foundational stance and methodological commitments, even if convergent support could be found with respect to understanding some of the aspects of consciousness and emergence).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00030/full#supplementary-material

REFERENCES

Aadithya, K. V., Ravindran, B., Michalak, T. P., and Jennings, N. R. (2010). "Efficient computation of the shapley value for centrality in networks," in *Internet and Network Economics Lecture Notes in Computer Science*, ed. A. Saberi (Springer Berlin Heidelberg), 1–13.

Abrego, L., and Zaikin, A. (2019). Integrated information as a measure of cognitive processes in coupled genetic repressilators. *Entropy* 21:382. doi: 10.3390/e21040382

Adams, R., Shipp, S., and Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Struct. Funct.* 218, 611–643. doi: 10.1007/s00429-012-0475-5

Albantakis, L. (2017). *A Tale of Two Animats: What does it take to have goals? ArXiv170510854 Cs Q-Bio*. Available online at: http://arxiv.org/abs/1705.10854 (accessed June 13, 2019).

Albantakis, L., and Tononi, G. (2019). Causal composition: structural differences among dynamically equivalent systems. *Entropy* 21:989. doi: 10.3390/e21100989

Arese Lucini, F., Del Ferraro, G., Sigman, M., and Makse, H. A. (2019). How the brain transitions from conscious to subliminal perception. *Neuroscience* 411, 280–290. doi: 10.1016/j.neuroscience.2019.03.047

Atasoy, S., Deco, G., Kringelbach, M. L., and Pearson, J. (2018). Harmonic brain modes: a unifying framework for linking space and time in brain dynamics. *Neuroscientist* 24, 277–293. doi: 10.1177/1073858417728032.

Atasoy, S., Donnelly, I., and Pearson, J. (2016). Human brain networks function in connectome-specific harmonic waves. *Nat. Commun.* 7:10340. doi: 10.1038/ncomms10340

Atasoy, S., Roseman, L., Kaelen, M., Kringelbach, M. L., Deco, G., and Carhart-Harris, R. L. (2017). Connectome-harmonic decomposition of human brain activity reveals dynamical repertoire re-organization under LSD. *Sci. Rep.* 7:17661. doi: 10.1038/s41598-017-17546-0

Baars, B. J. (1993). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Badcock, P. B., Friston, K. J., and Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Phys. Life Rev.* 31, 104–121. doi: 10.1016/j.plrev.2018.10.002

Bagherzadeh, Y., Baldauf, D., Pantazis, D., and Desimone, R. (2019). Alpha Synchrony and the Neurofeedback Control of Spatial Attention. *Neuron* 105, 577–587.e5. doi: 10.1016/j.neuron.2019.11.001

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron 95,* 709–721.e5. doi: 10.1016/j.neuron.2017.06.041

Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *J. Neurosci.* 38, 9689–9699. doi: 10.1523/JNEUROSCI.0251-18.2018

Balduzzi, D., and Tononi, G. (2009). Qualia: the geometry of integrated information. *PLoS Comput. Biol.* 5:e1000462. doi: 10.1371/journal.pcbi.1000462

Barrett, A. B., and Mediano, P. A. M. (2019). The phi measure of integrated information is not well-defined for general physical systems. *ArXiv190204321 Q-Bio*. Available online at: http://arxiv.org/abs/1902.04321 (accessed March 29, 2020).

Barsalou, L. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645. doi: 10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 1281–1289. doi: 10.1098/rstb.2008.0319

Barsalou, L. W. (2010). Grounded cognition: past, present, and future. *Top. Cogn. Sci.* 2, 716–724. doi: 10.1111/j.1756-8765.2010.01115.x

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J. R., et al. (2015). Visual areas exert feedforward and

feedback influences through distinct frequency channels. *Neuron* 85, 390–401. doi: 10.1016/j.neuron.2014.12.018

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neurosci. Conscious.* 2018: niy007. doi: 10.1093/nc/niy007

Beaty, R. E., Benedek, M., Barry Kaufman, S., and Silvia, P. J. (2015). Default and executive network coupling supports creative idea production. *Sci. Rep.* 5:10964. doi: 10.1038/srep10964

Beaty, R. E., Benedek, M., Wilkins, R. W., Jauk, E., Fink, A., Silvia, P. J., et al. (2014). Creativity and the default network: a functional connectivity analysis of the creative brain at rest. *Neuropsychologia* 64, 92–98. doi: 10.1016/j.neuropsychologia.2014.09.019

Beaty, R. E., Kenett, Y. N., Christensen, A. P., Rosenberg, M. D., Benedek, M., Chen, Q., et al. (2018). Robust prediction of individual creative ability from brain functional connectivity. *Proc. Natl. Acad. Sci. U. S. A.* 115, 1087–1092. doi: 10.1073/pnas.1713532115

Bejan, A. (2016). *The Physics of Life: The Evolution of Everything*. New York, NY: Macmillan Publishers.

Bengio, Y. (2017). The consciousness prior. *ArXiv170908568 Cs Stat*. Available online at: http://arxiv.org/abs/1709.08568 (accessed June 11, 2019).

Berrou, C., and Glavieux, A. (1996). Near optimum error correcting coding and decoding: turbo-codes. *IEEE Trans. Commun.* 44, 1261–1271. doi: 10.1109/26.539767

Betzel, R. F., Fukushima, M., He, Y., Zuo, X.-N., and Sporns, O. (2016). Dynamic fluctuations coincide with periods of high and low modularity in resting-state functional brain networks. *NeuroImage* 127, 287–297. doi: 10.1016/j.neuroimage.2015.12.001

Biehl, M., Pollock, F. A., and Kanai, R. (2020). *A Technical Critique of the Free Energy Principle as Presented in "Life as We Know it" and Related Works*. Available online at: https://arxiv.org/abs/2001.06408v2 (accessed March 28, 2020).

Block, N. (2008). Phenomenal and access consciousness ned block and cynthia macdonald: consciousness and cognitive access. *Proc. Aristot. Soc.* 108, 289–317. doi: 10.1111/j.1467-9264.2008.00247.x

Block, N. (2018). If perception is probabilistic, why does it not seem probabilistic? *Philos. Trans. R. Soc. B Biol. Sci.* 373:20170341. doi: 10.1098/rstb.2017.0341

Bogolyubov, N. N., and Shirkov, D. V. (1980). Introduction to the theory of quantized fields. *Intersci Monogr Phys Astron* 3, 1–720.

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., and Tononi, G. (2017). Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? *Clin. Neuroimag. Evidence. J. Neurosci.* 37, 9603–9613. doi: 10.1523/JNEUROSCI.3218-16.2017

Brewer, J. A., Garrison, K. A., and Whitfield-Gabrieli, S. (2013). What about the "Self" is processed in the posterior cingulate cortex? *Front. Hum. Neurosci.* 7:647. doi: 10.3389/fnhum.2013.00647

Brewer, J. A., Worhunsky, P. D., Gray, J. R., Tang, Y.-Y., Weber, J., and Kober, H. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. *Proc. Natl. Acad. Sci.U. S. A.* 108, 20254–9. doi: 10.1073/pnas.1112029108

Brillouin, L. (1951). Maxwell's demon cannot operate: information and entropy. *I. J. Appl. Phys.* 22, 334–337. doi: 10.1063/1.1699951

Brown, R., Lau, H., and LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends Cogn. Sci.* 23, 754–768. doi: 10.1016/j.tics.2019.06.009

Buchsbaum, D., Bridgers, S., Skolnick Weisberg, D., and Gopnik, A. (2012). The power of possibility: causal learning, counterfactual reasoning, and pretend play. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 2202–2212. doi: 10.1098/rstb.2012.0122

Buonomano, D. (2017). *Your Brain Is a Time Machine: The Neuroscience and Physics of Time*. New York, NY: WW Norton and Company.

Buzsáki, G., and Watson, B. O. (2012). Brain rhythms and neural syntax: implications for efficient coding of cognitive content and neuropsychiatric disease. *Dialogues Clin. Neurosci.* 14, 345–367.

Campbell, J. O. (2016). Universal darwinism as a process of bayesian inference. *Front. Syst. Neurosci.* 10:49. doi: 10.3389/fnsys.2016.00049

Canolty, R. T., and Knight, R. T. (2010). The functional role of cross-frequency coupling. *Trends Cogn. Sci.* 14, 506–515. doi: 10.1016/j.tics.2010.09.001

Carhart-Harris, R. L. (2018). The entropic brain - revisited. *Neuropharmacology* 142, 167–178. doi: 10.1016/j.neuropharm.2018.03.010

Carhart-Harris, R. L., and Friston, K. J. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain J. Neurol.* 133, 1265–1283. doi: 10.1093/brain/awq010

Carroll, S. (2016). *The Big Picture: On the Origins of Life, Meaning, and the Universe Itself*. New York, NY: Penguin Random House.

Chalmers, D. J. (1997). Moving forward on the problem of consciousness. *J. Conscious. Stud.* Available online at: http://cogprints.org/317/ (accessed June 11, 2019).

Chalmers, D. J. (2018). The meta-problem of consciousness. *J. Conscious. Stud.* 25, 6–61.

Chang, A. Y. C., Biehl, M., Yu, Y., and Kanai, R. (2019). *Information closure theory of consciousness. ArXiv190913045 Q-Bio*. Available online at: http://arxiv.org/abs/1909.13045 (accessed October 26, 2019).

Chao, Z. C., Takaura, K., Wang, L., Fujii, N., and Dehaene, S. (2018). Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. *Neuron* 100, 1252–1266.e3. doi: 10.1016/j.neuron.2018.10.004

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nat. Neurosci.* 20, 115–125. doi: 10.1038/nn.4450

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Clark, A. (2018). Beyond the "Bayesian Blur": predictive processing and the nature of subjective experience. *J. Conscious. Stud.* 25, 71–87. Available online at: https://www.ingentaconnect.com/content/imp/jcs/2018/00000025/f0020003/art00004

Cohen, J. R., and D'Esposito, M. (2016). The segregation and integration of distinct brain networks and their relationship to cognition. *J. Neurosci.* 36, 12083–12094. doi: 10.1523/JNEUROSCI.2965-15.2016

Conant, R. C., and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1, 89–97. doi: 10.1080/00207727008920220

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., Campbell, J. O., and Friston, K. J. (2018). A variational approach to niche construction. *J. R. Soc. Interface* 15:20170685. doi: 10.1098/rsif.2017.0685

Corcoran, A. W., Pezzulo, G., and Hohwy, J. (2020). From allostatic agents tocounterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biol Philos.* 35, 32. doi: 10.1007/s10539-020-09746-2.

Crick, F., and Koch, C. (2003). A framework for consciousness. *Nat. Neurosci.* 6, 119–126. doi: 10.1038/nn0203-119

Damasio, A. (2012). *Self Comes to Mind: Constructing the Conscious Brain*. *Reprint Edn*. New York: Vintage.

Davey, C. G., and Harrison, B. J. (2018). The brain's center of gravity: how the default mode network helps us to understand the self. *World Psychiatry* 17, 278–279. doi: 10.1002/wps.20553

Dawkins, R. (1999). *The Extended* Phenotype: *The Long Reach of the Gene*. Revised. Oxford University Press: USA.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The helmholtz machine. *Neural Comput.* 7, 889–904.

De Kock, L. (2016). Helmholtz's kant revisited (Once more). The all-pervasive nature of helmholtz's struggle with kant's anschauung. *Stud. Hist. Philos. Sci.* 56, 20–32. doi: 10.1016/j.shpsa.2015.10.009

Deacon, T. W. (2011). *Incomplete Nature: How Mind Emerged from Matter. 1st Edn*. New York, NY: WW Norton and Company.

Deco, G., and Kringelbach, M. L. (2016). Metastability and coherence: extending the communication through coherence hypothesis using a whole-brain computational perspective. *Trends Neurosci.* 39, 125–135. doi: 10.1016/j.tins.2016.01.001

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York, New York: Viking.

Dehaene, S. (2020). How We Learn: Why Brains Learn Better Than Any Machine... for Now. Penguin.

Dehaene, S., and Changeux, J.-P. (2005). Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. *PLOS Biol.* 3:e141. doi: 10.1371/journal.pbio.0030141

Dehaene, S., and Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227. doi: 10.1016/j.neuron.2011.03.018

Dennett, D. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. *1st Edn*. New York, NY: WW Norton and Company.

Dennett, D. C. (1992). "The self as a center of narrative gravity," in *Self and Consciousness: Multiple Perspectives*, eds F. Kessel, P. Cole, and D. Johnson (Hillsdale, NJ: Erlbaum).

Dennett, D. C. (2018). Facing up to the hard question of consciousness. *Philos. Trans. R. Soc. B Biol. Sci.* 373:20170342. doi: 10.1098/rstb.2017.0342

Edelman, G. (2004). *Wider than the Sky: The Phenomenal Gift of Consciousness*. *1st Edn*. New Haven, CT: Yale University Press.

Edelman, G., Gally, J. A., and Baars, B. J. (2011). Biology of consciousness. *Front. Psychol.* 2:4. doi: 10.3389/fpsyg.2011.00004

Edelman, G. J. (1987). *Neural Darwinism: The Theory Of Neuronal Group Selection*. *1st Edn*. New York, NY: Basic Books.

Elton, M. (2000). Consciouness: only at the personal level. *Philos. Explor.* 3, 25–42. doi: 10.1080/13869790008520979

Esfahlani, F. Z., Jo, Y., Faskowitz, J., Byrge, L., Kennedy, D., Sporns, O., et al. (2020). High-amplitude co-fluctuations in cortical activity drive functional connectivity. *bioRxiv*, 800045. doi: 10.1101/800045

Fallon, F. (2018). Integrated information theory, searle, and the arbitrariness question. *Rev. Phil. Psych.* 1–17. doi: 10.1007/s13164-018-0409-0

Feiten, T. E. (2020). Mind after uexküll: a foray into the worlds of ecological psychologists and enactivists. *Front. Psychol.* 11:480. doi: 10.3389/fpsyg.2020.00480

Fontenele, A. J., de Vasconcelos, N. A. P., Feliciano, T., Aguiar, L. A. A., Soares-Cunha, C., Coimbra, B., et al. (2019). Criticality between Cortical States. *Phys. Rev. Lett.* 122:208101. doi: 10.1103/PhysRevLett.122.208101

Fraccaro, M., Kamronn, S., Paquet, U., and Winther, O. (2017). *A disentangled recognition and nonlinear dynamics model for unsupervised learning. ArXiv171005741 Cs Stat*. Available online at: http://arxiv.org/abs/1710.05741 (accessed June 14, 2019).

Frank, S. A. (2012). Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. *J. Evol. Biol.* 25, 2377–2396. doi: 10.1111/jeb.12010

Fries, P. (2015). Rhythms for cognition: communication through coherence. *Neuron* 88, 220–235. doi: 10.1016/j.neuron.2015.09.034

Friston, K., Da Costa, L., and Parr, T. (2020a). *Some Interesting Observations on the Free Energy Principle*. Available online at: https://arxiv.org/abs/2002.04501v1 (accessed March 28, 2020).

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp.* 2, 56–78. doi: 10.1002/hbm.460020107

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J. (2013). Life as we know it. *J. R. Soc. Interface* 10:20130475. doi: 10.1098/rsif.2013.0475

Friston, K. J. (2017). Self-evidencing babies: commentary on "Mentalizing homeostasis: the social origins of interoceptive inference" by fotopoulou and tsakiris. *Neuropsychoanalysis* 19, 43–47. doi: 10.1080/15294145.2017.1295216

Friston, K. J. (2018). Am i self-conscious? (or does self-organization entail self-consciousness?). *Front. Psychol.* 9:579. doi: 10.3389/fpsyg.2018.00579

Friston, K. J. (2019). *A free energy principle for a particular physics. ArXiv190610184 Q-Bio*. Available online at: http://arxiv.org/abs/1906.10184 (accessed July 1, 2019).

Friston, K. J., Breakspear, M., and Deco, G. (2012a). Perception and self-organized instability. *Front. Comput. Neurosci.* 6:44. doi: 10.3389/fncom.2012.00044

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017a). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K. J., and Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex J. Devoted Study Nerv. Syst. Behav.* 68, 129–143. doi: 10.1016/j.cortex.2015.03.025

Friston, K. J., Kahan, J., Razi, A., Stephan, K. E., and Sporns, O. (2014). On nodes and modes in resting state fMRI. *NeuroImage* 99, 533–547. doi: 10.1016/j.neuroimage.2014.05.056

Friston, K. J., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Friston, K. J., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Friston, K. J., Levin, M., Sengupta, B., and Pezzulo, G. (2015). Knowing one's place: a free-energy approach to pattern regulation. *J. R. Soc. Interface* 12:20141383. doi: 10.1098/rsif.2014.1383

Friston, K. J., Parr, T., and de Vries, B. (2017b). The graphical brain: belief propagation and active inference. *Netw. Neurosci.* 1, 381–414. doi: 10.1162/NETN_a_00018

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017c). Deep temporal models and active inference. *Neurosci. Biobehav. Rev.* 77, 388–402. doi: 10.1016/j.neubiorev.2017.04.009

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., et al. (2012b). Dopamine, affordance and active inference. *PLoS Comput. Biol.* 8:e1002327. doi: 10.1371/journal.pcbi.1002327

Friston, K. J., Wiese, W., and Hobson, J. A. (2020b). Sentience and the origins of consciousness: from Cartesian duality to Markovian monism. *Entropy* 22:516. doi: 10.3390/e22050516

Fuster, J. M. (2009). Cortex and memory: emergence of a new paradigm. *J. Cogn. Neurosci.* 21, 2047–2072. doi: 10.1162/jocn.2009.21280

Gazzaniga, M. S. (2018). *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*. New York, NY: Farrar Straus and Giroux.

George, D., and Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Comput. Biol.* 5:e1000532. doi: 10.1371/journal.pcbi.1000532

Gramann, K., Onton, J., Riccobon, D., Mueller, H. J., Bardins, S., and Makeig, S. (2010). Human brain dynamics accompanying use of egocentric and allocentric reference frames during navigation. *J. Cogn. Neurosci.* 22, 2836–2849. doi: 10.1162/jocn.2009.21369

Graziano, M. S. A. (2013). *Consciousness and the Social Brain*. Oxford: Oxford University Press.

Graziano, M. S. A. (2019). *Rethinking Consciousness: a Scientific Theory of Subjective Experience*. *1st Edn*. New York: WW Norton and Company.

Gross, S. (2018). Perceptual consciousness and cognitive access from the perspective of capacity-unlimited working memory. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 373:20170343. doi: 10.1098/rstb.2017.0343

Grossberg, S. (2017). Towards solving the hard problem of consciousness: the varieties of brain resonances and the conscious experiences that they support. *Neural Netw.* 87, 38–95. doi: 10.1016/j.neunet.2016.11.003

Guterstam, A., Björnsdotter, M., Gentile, G., and Ehrsson, H. H. (2015). Posterior cingulate cortex integrates the senses of self-location and body ownership. *Curr. Biol.* 25, 1416–1425. doi: 10.1016/j.cub.2015.03.059

Ha, D., and Schmidhuber, J. (2018). World models. *ArXiv180310122 Cs Stat*. doi: 10.5281/zenodo.1207631

Hahn, G., Ponce-Alvarez, A., Deco, G., Aertsen, A., and Kumar, A. (2019). Portraits of communication in neuronal networks. *Nat. Rev. Neurosci.* 20, 117–127. doi: 10.1038/s41583-018-0094-0

Haimovici, A., Tagliazucchi, E., Balenzuela, P., and Chialvo, D. R. (2013). Brain organization into resting state networks emerges at criticality on a model of the human connectome. *Phys. Rev. Lett.* 110:178101. doi: 10.1103/PhysRevLett.110.178101

Haken, H. (1977). Synergetics. *Phys. Bull.* 28:412.

Haken, H. (1992). "Synergetics of the brain: an outline of some basic ideas," in *Induced Rhythms in the Brain Brain Dynamics*, eds. E. Başar and T. H. Bullock (Boston, MA: Birkhäuser Boston), 417–421. doi: 10.1007/978-1-4757-1281-0_23

Harper, M. (2011). Escort evolutionary game theory. *Phys. Nonlinear Phenom.* 240, 1411–1415. doi: 10.1016/j.physd.2011.04.008

Harrison, C. W. (1952). Experiments with linear prediction in television. *Bell Syst. Tech. J.* 31, 764–783. doi: 10.1002/j.1538-7305.1952.tb01405.x

Hassabis, D., and Maguire, E. A. (2009). The construction system of the brain. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 364, 1263–1271. doi: 10.1098/rstb.2008.0296

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., and Schacter, D. L. (2014). Imagine all the people: how the brain creates and uses personality models to predict behavior. *Cereb. Cortex* 24, 1979–1987. doi: 10.1093/cercor/bht042

Hattori, R., Danskin, B., Babic, Z., Mlynaryk, N., and Komiyama, T. (2019). Area-specificity and plasticity of history-dependent value coding during learning. *Cell 177*, 1858–1872.e15. doi: 10.1016/j.cell.2019.04.027

Haun, A., and Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy* 21:1160. doi: 10.3390/e21121160

Hawkins, J., and Blakeslee, S. (2004). *On Intelligence*. Adapted. New York, NY: Times Books.

Hayek, F. A. (1952). *The Sensory Order: An Inquiry into the Foundations of Theoretical Psychology*. Chicago, IL: University Of Chicago Press.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory. New Edn*. East Sussex: Psychology Press.

Helmholtz, H. (1878). *"The Facts in Perception," in Selected Writings of Hermann Helmholtz*, ed. R. Kahl (Wesleyan University Press).

Heuvel, M. P., van den, Kahn, R. S., Goñi, J., and Sporns, O. (2012). High-cost, high-capacity backbone for global brain communication. *Proc. Natl. Acad. Sci. U. S. A.* 109, 11372–11377. doi: 10.1073/pnas.1203593109

Hinton, G. (2017). How to do Backpropagation in a Brain. 22.

Hirsh, J. B., Mar, R. A., and Peterson, J. B. (2013). Personal narratives as the highest level of cognitive integration. *Behav. Brain Sci.* 36, 216–217. doi: 10.1017/S0140525X12002269

Hobson, J. A., and Friston, K. J. (2016). A response to our theatre critics. *J. Conscious. Stud.* 23, 245–254. Available online at: https://www.ingentaconnect.com/content/imp/jcs/2016/00000023/f0020003/art00012

Hoel, E. P., Albantakis, L., Marshall, W., and Tononi, G. (2016). Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neurosci. Conscious.* 2016: niw012. doi: 10.1093/nc/niw012

Hoffman, D. D., and Singh, M. (2012). Computational evolutionary perception. *Perception* 41, 1073–1091. doi: 10.1068/p7275

Hoffmann, H., and Payton, D. W. (2018). Optimization by self-organized criticality. *Sci. Rep.* 8:2358. doi: 10.1038/s41598-018-20275-7

Hofstadter, D. R., and Mitchell, M. (1994). "*The copycat* project: a model of mental fluidity and analogy-making," In *Advances in Connectionist and Neural Computation Theory, Vol. 2. Analogical connections,* eds. K. J. Holyoak and J. A. Barnden (New York, NY: Ablex Publishing), 31–112.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Front. Psychol.* 3:96. doi: 10.3389/fpsyg.2012.00096

Hohwy, J. (2013). *Perceptual Unity in Action*. Oxford University Press Available online at: https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199682737.001.0001/acprof-9780199682737-chapter-11 (accessed June 15, 2019).

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259–285. doi: 10.1111/nous.12062

Hohwy, J. (2017). *"How to entrain your evil demon," in Philosophy and Predictive Processing*, eds. T. Metzinger and W. Wiese (Mainz: MIND Group), 1–15.

Hohwy, J. (2020). Self-supervision, normativity and the free energy principle. *Synthese* doi: 10.1007/s11229-020-02622-2

Hordijk, W., and Steel, M. (2015). Autocatalytic sets and boundaries. *J. Syst. Chem.* 6:1. doi: 10.1186/s13322-014-0006-2

Huang, Z., Zhang, J., Wu, J., Mashour, G. A., and Hudetz, A. G. (2020). Temporal circuit of macroscale dynamic brain activity supports human consciousness. *Sci. Adv*. 6:eaaz0087. doi: 10.1126/sciadv.aaz0087

Humphrey, N. (2017). The invention of consciousness. *Topoi* 39, 13–21. doi: 10.1007/s11245-017-9498-0

Jafri, H. H., Singh, R. K. B., and Ramaswamy, R. (2016). Generalized synchrony of coupled stochastic processes with multiplicative noise. *Phys. Rev. E* 94:052216. doi: 10.1103/PhysRevE.94.052216

Jann, K., Dierks, T., Boesch, C., Kottlow, M., Strik, W., and Koenig, T. (2009). BOLD correlates of EEG alpha phase-locking and the fMRI default mode network. *NeuroImage* 45, 903–916. doi.org/10.1016/j.neuroimage.2009.01.001

Joslyn, C. (2000). Levels of control and closure in complex semiotic systems. *Ann. NY. Acad. Sci.* 901, 67–74. doi: 10.1111/j.1749-6632.2000.tb06266.x

Kachman, T., Owen, J. A., and England, J. L. (2017). Self-organized resonance during search of a diverse chemical space. *Phys. Rev. Lett.* 119:038001. doi: 10.1103/PhysRevLett.119.038001

Kaila, V., and Annila, A. (2008). Natural selection for least action. *Proc. R. Soc. Math. Phys. Eng. Sci.* 464, 3055–3070. doi: 10.1098/rspa.2008.0178

Kanai, R., Chang, A., Yu, Y., Magrans de Abril, I., Biehl, M., and Guttenberg, N. (2019). Information generation as a functional basis of consciousness. *Neurosci. Conscious.* 2019:niz016. doi: 10.1093/nc/niz016

Kant, I. (1781). *Critique of Pure Reason*. eds. P. Guyer and A. W. Wood Cambridge: Cambridge University Press.

Kauffman, S., and Clayton, P. (2006). On emergence, agency, and organization. *Biol. Philos.* 21, 501–521. doi: 10.1007/s10539-005-9003-9

Kenett, Y. N., Medaglia, J. D., Beaty, R. E., Chen, Q., Betzel, R. F., Thompson-Schill, S. L., et al. (2018). Driving the brain towards creativity and intelligence: a network control theory analysis. *Neuropsychologia* 118, 79–90. doi: 10.1016/j.neuropsychologia.2018.01.001

Kerr, C. E., Sacchet, M. D., Lazar, S. W., Moore, C. I., and Jones, S. R. (2013). Mindfulness starts with the body: somatosensory attention and top-down modulation of cortical alpha rhythms in mindfulness meditation. *Front. Hum. Neurosci.* 7:12. doi: 10.3389/fnhum.2013.00012

Khajehabdollahi, S., Abeyasinghe, P. M., Owen, A. M., and Soddu, A. (2019). The emergence of integrated information, complexity, and consciousness at criticality. *bioRxiv* 521567. doi: 10.1101/521567

Kingma, D. P., and Welling, M. (2014). Auto-Encoding Variational Bayes. *ArXiv13126114 Cs Stat*. Available online at: http://arxiv.org/abs/1312.6114 (accessed March 29, 2020).

Kirchhoff, M., Parr, T., Palacios, E., Friston, K. J., and Kiverstein, J. (2018). The markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface* 15:20170792. doi: 10.1098/rsif.2017.0792

Knyazev, G. G., Slobodskoj-Plusnin, J. Y., Bocharov, A. V., and Pylkova, L. V. (2011). The default mode network and EEG alpha oscillations: An independent component analysis. *Brain Res.* 1402, 67–79. doi: 10.1016/j.brainres.2011.05.052

Koch, C. (2012). *Consciousness: Confessions of a Romantic Reductionist*. Cambridge, MA: MIT Press.

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Kosiorek, A., Sabour, S., Teh, Y. W., and Hinton, G. E. (2019). "Stacked capsule autoencoders," in *Advances in Neural Information Processing Systems eds* H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc.), 15512–15522. Available online at: http://papers.nips.cc/paper/9684-stacked-capsule-autoencoders.pdf (accessed May 14, 2020).

Koster, R., Chadwick, M. J., Chen, Y., Berron, D., Banino, A., Düzel, E., et al. (2018). Big-loop recurrence within the hippocampal system supports integration of information across episodes. *Neuron* 99, 1342–1354.e6. doi: 10.1016/j.neuron.2018.08.009

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). Revealing the Dark Secrets of BERT. *ArXiv190808593 Cs Stat*. Available online at: http://arxiv.org/abs/1908.08593 (accessed October 13, 2019).

Krotov, D., Dubuis, J. O., Gregor, T., and Bialek, W. (2014). Morphogenesis at criticality. *Proc. Natl. Acad. Sci. U. S. A.* 111, 3683–3688. doi: 10.1073/pnas.1324186111

Lahav, N., Sendiña-Nadal, I., Hens, C., Ksherim, B., Barzel, B., Cohen, R., et al. (2018). Synchronization of chaotic systems: a microscopic description. *Phys. Rev. E* 98:052204. doi: 10.1103/PhysRevE.98.052204

Lakoff, G. (2014). Mapping the brain's metaphor circuitry: metaphorical thought in everyday reason. *Front. Hum. Neurosci.* 8:958. doi: 10.3389/fnhum.2014.00958

Laland, N., Uller, T., Feldman Marcus, W., Sterelny, K., Müller Gerd, B., Moczek, A., et al. (2015). The extended evolutionary synthesis: its structure, assumptions and predictions. *Proc. R. Soc. B Biol. Sci.* 282:20151019. doi: 10.1098/rspb.2015.1019

Lane, N. (2016). *The Vital Question: Why is Life the Way it Is?* London: Profile Books.

Lau, H., and Michel, M. (2019). On the dangers of conflating strong and weak versions of a theory of consciousness. *PsyArXiv* doi: 10.31234/osf.io/hjp3s

LeDoux, J. (2019). *The Deep History of Ourselves: The Four-Billion-Year Story of How We Got Conscious Brains*. New York, NY: Viking.

Leibniz, G. W. (1714). *Monadologie; Trans. R. Ariew and D. Garber as Monadology in Leibniz: Philosophical Essays.* Indianapolis, IN; Cambridge, MA: Hackett Publishing Company. 1989.

Li, M., Woelfer, M., Colic, L., Safron, A., Chang, C., Heinze, H.-J., et al. (2018). Default mode network connectivity change corresponds to ketamine's delayed glutamatergic effects. *Eur. Arch. Psychiatry Clin. Neurosci.* 270, 207–216. doi: 10.1007/s00406-018-0942-y

Li, S.-H., and Wang, L. (2018). Neural network renormalization group. *Phys. Rev. Lett.* 121:260601. doi: 10.1103/PhysRevLett.121.260601

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. Nat. *Rev. Neurosci.* 1–12. doi: 10.1038/s41583-020-0277-3

Lin, H. W., and Tegmark, M. (2017). Criticality in formal languages and statistical physics. *Entropy* 19:299. doi: 10.3390/e19070299

Lin, H. W., Tegmark, M., and Rolnick, D. (2017). Why does deep and cheap learning work so well? *J. Stat. Phys.* 168, 1223–1247. doi: 10.1007/s10955-017-1836-5

Linson, A., Clark, A., Ramamoorthy, S., and Friston, K. J. (2018). The active inference approach to ecological perception: general information dynamics for natural and artificial embodied cognition. *Front. Robot. AI* 5:21. doi: 10.3389/frobt.2018.00021

Liu, J., Kumar, A., Ba, J., Kiros, J., and Swersky, K. (2019). Graph normalizing flows. *arXiv [Pre-print]. arXiv*:1905.13177. Available online at: http://arxiv.org/abs/1905.13177 (accessed May 24, 2020).

Lycan, W. G. (1996). *Consciousness and Experience*. Cambridge, MA: MIT Press.

MacKay, D. G. (2019). "*Remembering: what 50 years of research with famous amnesia patient HM*," in *Can Teach Us about Memory and How It Works*. Amherst, NY: Prometheus Books.

Madl, T., Baars, B. J., and Franklin, S. (2011). The timing of the cognitive cycle. *PloS ONE* 6:e14803. doi: 10.1371/journal.pone.0014803

Mark, J. T., Marion, B. B., and Hoffman, D. D. (2010). Natural selection and veridical perceptions. *J. Theor. Biol.* 266, 504–515. doi: 10.1016/j.jtbi.2010.07.020

Markram, H., Gerstner, W., and Sjöström, P. J. (2011). A history of spike-timing-dependent plasticity. *Front. Synaptic Neurosci.* 3:4. doi: 10.3389/fnsyn.2011.00004

Marr, D. (1983). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: Henry Holt and Company.

Marshall, W., Gomez-Ramirez, J., and Tononi, G. (2016). Integrated information and state differentiation. *Front. Psychol.* 7:926. doi: 10.3389/fpsyg.2016.00926

Marshall, W., Kim, H., Walker, S. I., Tononi, G., and Albantakis, L. (2017). How causal analysis can reveal autonomy in models of biological systems. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 375:20160358. doi: 10.1098/rsta.2016.0358

Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798. doi: 10.1016/j.neuron.2020.01.026

Maturana, H. R., and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living. 1st Edn*. Dordrecht, Holland; Boston: D. Reidel Publishing Company.

Mediano, P. A. M., Rosas, F., Carhart-Harris, R. L., Seth, A. K., and Barrett, A. B. (2019a). Beyond integrated information: a taxonomy of information dynamics phenomena. *ArXiv190902297 Phys. Q-Bio*. Available online at: http://arxiv.org/abs/1909.02297 (accessed November 23, 2019).

Mediano, P. A. M., Seth, A. K., and Barrett, A. B. (2019b). Measuring integrated information: comparison of candidate measures in theory and simulation. *Entropy* 21:17. doi: 10.3390/e21010017

Metzinger, T. (2010). *The Ego Tunnel*: *The Science of the Mind and the Myth of the Self*. New York, NY: Basic Books.

Meunier, D., Lambiotte, R., and Bullmore, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Front. Neurosci.* 4:200. doi: 10.3389/fnins.2010.00200

Michalareas, G., Vezoli, J., van Pelt, S., Schoffelen, J.-M., Kennedy, H., and Fries, P. (2016). Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. *Neuron* 89, 384–397. doi: 10.1016/j.neuron.2015.12.018

Milliere, R., and Metzinger, T. (2020). Radical disruptions of self-consciousness. *Philos. Mind Sci.* 1, 1–1. doi: 10.33735/phimisci.2020.I.50

Mišić, B., Betzel, R. F., Nematzadeh, A., Goñi, J., Griffa, A., Hagmann, P., et al. (2015). Cooperative and competitive spreading dynamics on the human connectome. *Neuron* 86, 1518–1529. doi: 10.1016/j.neuron.2015.05.035

Mohr, H., Wolfensteller, U., Betzel, R. F., Miši,ć, B., Sporns, O., Richiardi, J., et al. (2016). Integration and segregation of large-scale brain networks during short-term task automatization. *Nat. Commun.* 7:13217. doi: 10.1038/ncomms13217

Muller, L., Chavane, F., Reynolds, J., and Sejnowski, T. J. (2018). Cortical travelling waves: mechanisms and computational principles. *Nat. Rev. Neurosci.* 19, 255–268. doi: 10.1038/nrn.2018.20

Mumford, D. (1991). On the computational architecture of the neocortex. *Biol. Cybern.* 65, 135–145. doi: 10.1007/BF00202389

Nagel, T. (1974). What is it like to be a bat? *Philos. Rev.* 83, 435–450. doi: 10.2307/2183914

Northoff, G. (2012). Immanuel kant's mind and the brain's resting state. *Trends Cogn. Sci.* 16, 356–359. doi: 10.1016/j.tics.2012.06.001

Northoff, G., and Huang, Z. (2017). How do the brain's time and space mediate consciousness and its different dimensions? *Temporo-spatial theory of consciousness (TTC). Neurosci. Biobehav. Rev.* 80, 630–645. doi: 10.1016/j.neubiorev.2017.07.013

Ódor, G., Dickman, R., and Ódor, G. (2015). Griffiths phases and localization in hierarchical modular networks. *Sci. Rep.* 5:14451. doi: 10.1038/srep14451

O'Regan, J. K., and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* 24, 939–973. doi: 10.1017/s0140525x01000115

O'Reilly, R. C., Wyatte, D. R., and Rohrlich, J. (2017). Deep predictive learning: a comprehensive model of three visual streams. *arXiv [Pre-print]. arXiv*:1709.04654. Available online at: http://arxiv.org/abs/1709.04654

Palacios, E. R., Isomura, T., Parr, T., and Friston, K. J. (2019). The emergence of synchrony in networks of mutually inferring neurons. *Sci. Rep.* 9:6412. doi: 10.1038/s41598-019-42821-7

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., and Friston, K. J. (2020). On markov blankets and hierarchical self-organisation. *J. Theor. Biol.* 486:110089. doi: 10.1016/j.jtbi.2019.110089

Palmer, S. E., Marre, O., Berry, M. J., and Bialek, W. (2015). Predictive information in a sensory population. *Proc. Natl. Acad. Sci. U. S. A.* 112, 6908–6913. doi: 10.1073/pnas.1506855112

Palva, S., and Palva, J. M. (2011). Functional roles of alpha-band phase synchronization in local and large-scale cortical networks. *Front. Psychol.* 2:204. doi: 10.3389/fpsyg.2011.00204

Papez, J. W. (1937). A proposed mechanism of emotion. *Arch. Neurol. Psychiatry* 38, 725–743. doi: 10.1001/archneurpsyc.1937.02260220069003

Parr, T., and Friston, K. J. (2018a). The anatomy of inference: generative models and brain structure. *Front. Comput. Neurosci.* 12:90. doi: 10.3389/fncom.2018.00090

Parr, T., and Friston, K. J. (2018b). The discrete and continuous brain: from decisions to movement-and back again. *Neural Comput.* 30, 2319–2347. doi: 10.1162/neco_a_01102

Parr, T., Markovic, D., Kiebel, S. J., and Friston, K. J. (2019). Neuronal message passing using mean-field, bethe, and marginal approximations. *Sci. Rep.* 9:1889. doi: 10.1038/s41598-018-38246-3

Pattee, H. H. (2001). The physics of symbols: bridging the epistemic cut. *Biosystems* 60, 5–21. doi: 10.1016/s0303-2647(01)00104-6

Payne, J. L., and Wagner, A. (2019). The causes of evolvability and their evolution. *Nat. Rev. Genet.* 20, 24–38. doi: 10.1038/s41576-018-0069-z

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco, CA: Elsevier. doi: 10.1016/C2009-0-27609-4

Pearl, J., and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books.

Petkova, M. D., Tkačik, G., Bialek, W., Wieschaus, E. F., and Gregor, T. (2019). Optimal decoding of cellular identities in a genetic network. *Cell 176*, 844–855.e15. doi: 10.1016/j.cell.2019.01.007

Pfeifer, R., and Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*. Cambridge, Mass: A Bradford Book.

Pletzer, B., Kerschbaum, H., and Klimesch, W. (2010). When frequencies never synchronize: the golden mean and the resting EEG. *Brain Res.* 1335, 91–102. doi: 10.1016/j.brainres.2010.03.074

Prinz, J. (2017). "The intermediate level theory of consciousness," in *The Blackwell Companion to Consciousness,* eds S. Schneider and M. Velmans (John Wiley and Sons, Ltd), 257–271. doi: 10.1002/9781119132363.ch18

Ramstead, M. J. D., Badcock, P. B., and Friston, K. J. (2018). Answering schrödinger's question: a free-energy formulation. *Phys. Life Rev.* 24, 1–16. doi: 10.1016/j.plrev.2017.09.001

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., et al. (2016). Cultural group selection plays an essential role in explaining human cooperation: a sketch of the evidence. *Behav. Brain Sci.* 39:e30. doi: 10.1017/S0140525X1400106X

Rudrauf, D., Bennequin, D., Granic, I., Landini, G., Friston, K. J., and Williford, K. (2017). A mathematical model of embodied consciousness. *J. Theor. Biol.* 428, 106–131. doi: 10.1016/j.jtbi.2017.05.032

Rudrauf, D., Lutz, A., Cosmelli, D., Lachaux, J.-P., and Le Van Quyen, M. (2003). From autopoiesis to neurophenomenology: Francisco varela's exploration of the biophysics of being. *Biol. Res.* 36, 27–65. doi: 10.4067/s0716-97602003000100005

Russell, S. J., and Subramanian, D. (1995). Provably bounded-optimal agents. *arXiv:cs/9505103*. Available online at: http://arxiv.org/abs/cs/9505103 (accessed October 26, 2019).

Safron, A. (2019a). Integrated world modeling theory (IWMT) revisited. *PsyArXiv* doi: 10.31234/osf.io/kjngh

Safron, A. (2019b). Multilevel evolutionary developmental optimization (MEDO): A theoretical framework for understanding preferences and selection dynamics. *ArXiv191013443 Econ Q-Bio Q-Fin*. Available online at: http://arxiv.org/abs/1910.13443 (accessed November 14, 2019).

Safron, A. (2019c). The radically embodied conscious cybernetic Bayesian brain: towards explaining the emergence of agency. doi: 10.31234/osf.io/udc42

Salehipour, H., Peltier, W. R., and Caulfield, C. P. (2018). Self-organized criticality of turbulence in strongly stratified mixing layers. *J. Fluid Mech.* 856, 228–256. doi: 10.1017/jfm.2018.695

Sato, J., Mossad, S. I., Wong, S. M., Hunt, B. A. E., Dunkley, B. T., Smith, M. L., et al. (2018). Alpha keeps it together: alpha oscillatory synchrony underlies working memory maintenance in young children. *Dev. Cogn. Neurosci.* 34, 114–123. doi: 10.1016/j.dcn.2018.09.001

Schartner, M. M., Carhart-Harris, R. L., Barrett, A. B., Seth, A. K., and Muthukumaraswamy, S. D. (2017). Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Sci. Rep.* 7:46421. doi: 10.1038/srep46421

Scheeringa, R., and Fries, P. (2019). Cortical layers, rhythms and BOLD signals. *NeuroImage* 197, 689–698. doi: 10.1016/j.neuroimage.2017.11.002

Schrödinger, E. (1926). An undulatory theory of the mechanics of atoms and molecules. *Phys. Rev.* 28, 1049–1070. doi: 10.1103/PhysRev.28.1049

Schrodinger, E. (1944). *What is Life?: With Mind and Matter and Autobiographical Sketches*. *Reprint Edn*. Cambridge; New York: Cambridge University Press.

Sedley, W., Gander, P. E., Kumar, S., Kovach, C. K., Oya, H., Kawasaki, H., et al. (2016). Neural signatures of perceptual inference. *Elife* 5:e11476. doi: 10.7554/eLife.11476

Sengupta, B., Tozzi, A., Cooray, G. K., Douglas, P. K., and Friston, K. J. (2016). Towards a neuronal gauge theory. *PLoS Biol.* 14:e1002400. doi: 10.1371/journal.pbio.1002400

Seth, A. K. (2015). "The cybernetic Bayesian brain: from interoceptive inference to sensorimotor contingencies," in *Open MIND*, eds J. M. Windt, and T. Metzinger (Frankfurt: MIND Group), 9–24.

Seth, A. K. (2016). *The hard problem of consciousness is a distraction from the real one – Anil K Seth | Aeon Essays*. Aeon. Available online at: https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one (accessed November 25, 2019).

Seth, A. K., and Tsakiris, M. (2018). Being a beast machine: the somatic basis of selfhood. *Trends Cogn. Sci.* 22, 969–981. doi: 10.1016/j.tics.2018.08.008

Shanahan, M. (2012). The brain's connective core and its role in animal cognition. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 2704–2714. doi: 10.1098/rstb.2012.0128

Shanahan, M., and Baars, B. (2005). Applying global workspace theory to the frame problem. *Cognition* 98, 157–176. doi: 10.1016/j.cognition.2004.11.007

Shea, N., and Frith, C. D. (2019). The global workspace needs metacognition. *Trends Cogn. Sci.* 23, 560–571. doi: 10.1016/j.tics.2019.04.007

Shine, J. M., Breakspear, M., Bell, P. T., Martens, K. A. E., Shine, R., Koyejo, O., et al. (2019). Human cognition involves the dynamic integration of neural activity and neuromodulatory systems. *Nat. Neurosci.* 22, 289–296. doi: 10.1038/s41593-018-0312-0

Singer, W. (2001). Consciousness and the binding problem. *Ann. N. Y. Acad. Sci.* 929, 123–146. doi: 10.1111/j.1749-6632.2001.tb05712.x

Smigielski, L., Scheidegger, M., Kometer, M., and Vollenweider, F. X. (2019). Psilocybin-assisted mindfulness training modulates self-consciousness and brain default mode network connectivity with lasting effects. *NeuroImage* 196, 207–215. doi: 10.1016/j.neuroimage.2019.04.009

Sormaz, M., Murphy, C., Wang, H., Hymers, M., Karapanagiotidis, T., Poerio, G., et al. (2018). Default mode network can support the level of detail in experience during active task states. *Proc. Natl. Acad. Sci. U. S. A.* 115, 9318–9323. doi: 10.1073/pnas.1721259115

Spelke, E. S., and Kinzler, K. D. (2007). Core knowledge. *Dev. Sci.* 10, 89–96. doi: 10.1111/j.1467-7687.2007.00569.x

Sporns, O. (2013). Network attributes for segregation and integration in the human brain. *Curr. Opin. Neurobiol.* 23, 162–171. doi: 10.1016/j.conb.2012.11.015

Stepp, N., and Turvey, M. T. (2010). On strong anticipation. *Cogn. Syst. Res.* 11, 148–164. doi: 10.1016/j.cogsys.2009.03.003

Steppa, C., and Holch, T. L. (2019). HexagDLy—Processing hexagonally sampled data with CNNs in PyTorch. *SoftwareX* 9, 193–198. doi: 10.1016/j.softx.2019.02.010

Strogatz, S. H. (2012). *Sync: How Order Emerges from Chaos In the Universe, Nature, and Daily Life*. New York, NY: Hachette Books.

Tagliazucchi, E., Carhart-Harris, R., Leech, R., Nutt, D., and Chialvo, D. R. (2014). Enhanced repertoire of brain dynamical states during the psychedelic experience. *Hum. Brain Mapp.* 35, 5442–5456. doi: 10.1002/hbm.22562

Takagi, K. (2018). Information-based principle induces small-world topology and self-organized criticality in a large scale brain network. *Front. Comput. Neurosci.* 12:65. doi: 10.3389/fncom.2018.00065

Tani, J. (2016). *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. New York, NY: Oxford University Press.

Tegmark, M. (2014). *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. New York, NY: Knopf Doubleday Publishing Group.

Tegmark, M. (2016). Improved measures of integrated information. *PLoS Comput. Biol.* 12:5123. doi: 10.1371/journal.pcbi.1005123

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci.* 5:42. doi: 10.1186/1471-2202-5-42

Tononi, G. (2013). "On the irreducibility of consciousness and its relevance to free will," in *Is Science Compatible with Free Will?*, eds A. Suarez and P. Adams (New York, NY: Springer), 147–176.

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* 17:450. doi: 10.1038/nrn.2016.44

Tononi, G., and Koch, C. (2015). Consciousness: here, there and everywhere? *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140167. doi: 10.1098/rstb.2014.0167

Touboul, J., and Destexhe, A. (2017). Power-law statistics and universal scaling in the absence of criticality. *Phys. Rev. E* 95:012413. doi: 10.1103/PhysRevE.95.012413

Towlson, E. K., Vértes, P. E., Ahnert, S. E., Schafer, W. R., and Bullmore, E. T. (2013). The rich club of the *C. elegans* neuronal connectome. *J. Neurosci.* 33, 6380–6387. doi: 10.1523/JNEUROSCI.3784-12.2013

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 237, 37–72.

Ullmann, D. (2007). Life and work of E.F.F. *Chladni. Eur. Phys. J. Spec. Top.* 145, 25–32. doi: 10.1140/epjst/e2007-00145-4

van den Heuvel, M. P., and Sporns, O. (2011). Rich-club organization of the human connectome. *J. Neurosci.* 31, 15775–15786. doi: 10.1523/JNEUROSCI.3539-11.2011

Varela, F., Lachaux, J. P., Rodriguez, E., and Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* 2, 229–239. doi: 10.1038/35067550

Vázquez-Rodríguez, B., Avena-Koenigsberger, A., Sporns, O., Griffa, A., Hagmann, P., and Larralde, H. (2017). Stochastic resonance at criticality in a network model of the human cortex. *Sci. Rep.* 7:13020. doi: 10.1038/s41598-017-13400-5

Viol, A., Palhano-Fontes, F., Onias, H., Araujo, D. B., de, and Viswanathan, G. M. (2017). Shannon entropy of brain functional complex networks under the influence of the psychedelic Ayahuasca. *Sci. Rep.* 7:7388. doi: 10.1038/s41598-017-06854-0

von Uexküll, J. (1957). A stroll through the worlds of animals and men. in *Instinctive Behavior: The Development of a Modern Concept*, ed C. H. Schiller (New York, NY: International Universities Press), 5–80.

Vul, E., Goodman, N., Griffiths, T. L., and Tenenbaum, J. B. (2014). One and done? *Optimal decisions from very few samples. Cogn. Sci.* 38, 599–637. doi: 10.1111/cogs.12101

Wang, S.-J., Hilgetag, C., and Zhou, C. (2011). Sustained activity in hierarchical modular neural networks: self-organized criticality and oscillations. *Front. Comput. Neurosci.* 5:30. doi: 10.3389/fncom.2011.00030

Wens, V., Bourguignon, M., Vander Ghinst, M., Mary, A., Marty, B., Coquelet, N., et al. (2019). Synchrony, metastability, dynamic integration, and competition in the spontaneous functional connectivity of the human brain. *NeuroImage*. 199, 313–324. doi: 10.1016/j.neuroimage.2019.05.081

Whittington, J. C. R., Muller, T. H., Mark, S., Barry, C., and Behrens, T. E. J. (2018). Generalisation of structural knowledge in the hippocampal-entorhinal system. *ArXiv180509042 Cs Q-Bio Stat*. Available online at: http://arxiv.org/abs/1805.09042 (accessed June 13, 2019).

Whyte, C. J., and Smith, R. (2020). The predictive global neuronal workspace: a formal active inference model of visual consciousness. *bioRxiv*.

Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenol. Cogn. Sci.* 16, 715–736. doi: 10.1007/s11097-016-9472-0

Williford, K., Bennequin, D., Friston, K., and Rudrauf, D. (2018). The projective consciousness model and phenomenal selfhood. *Front. Psychol.* 9:2571. doi: 10.3389/fpsyg.2018.02571

Wu, Y., Wayne, G., Graves, A., and Lillicrap, T. (2018). The kanerva machine: a generative distributed memory. *ArXiv180401756 Cs Stat*. Available online at: http://arxiv.org/abs/1804.01756 (accessed June 15, 2019).

Yufik, Y. M., and Friston, K. J. (2016). Life and understanding: the origins of "Understanding" in self-organizing nervous systems. *Front. Syst. Neurosci.* 10:98. doi: 10.3389/fnsys.2016.00098

Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit memories to other brains: constructing shared neural representations via communication. *Cereb. Cortex N Y. N* 1991 27, 4988–5000. doi: 10.1093/cercor/bhx202

Zhang, H., Watrous, A. J., Patel, A., and Jacobs, J. (2018). Theta and alpha oscillations are traveling waves in the human neocortex. *Neuron 98*, 1269–1281.e4. doi: 10.1016/j.neuron.2018.05.019

Ziporyn, B. (2004). *Being and Ambiguity: Philosophical Experiments with Tiantai Buddhism. 1st Edn*. Chicago, IL: Open Court.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

HYPOTHESIS AND THEORY
published: 14 August 2020
doi: 10.3389/fdata.2020.00027

Embodied Predictions, Agency, and Psychosis

*Pantelis Leptourgos and Philip R. Corlett**

Department of Psychiatry, Connecticut Mental Health Center, Yale University, New Haven, CT, United States

Psychotic symptoms, i.e., hallucinations and delusions, involve gross departures from conscious apprehension of consensual reality; respectively, perceiving and believing things that, according to same culture peers, do not obtain. In schizophrenia, those experiences are often related to abnormal sense of control over one's own actions, often expressed as a distorted sense of agency (i.e., passivity symptoms). Cognitive and computational neuroscience have furnished an account of these experiences and beliefs in terms of the brain's generative model of the world, which underwrites inferences to the best explanation of current and future states, in order to behave adaptively. Inference then involves a reliability-based trade off of predictions and prediction errors, and psychotic symptoms may arise as departures from this inference process, either an over- or under-weighting of priors relative to prediction errors. Surprisingly, there is empirical evidence in favor of both positions. Relatedly, there is evidence for both an enhanced and a diminished sense of agency in schizophrenia. How can this be? We argue that there is more than one generative model in the brain, and that ego- and allo-centric models operate in tandem. In brief, ego-centric models implement corollary discharge signals that cancel out the effects of self-generated actions while allo-centric models compare several hypothesis regarding the causes of sensory inputs (including the self among the potential causes). The two parallel hierarchies give rise to different levels of agency, with ego-centric models subserving "feelings of agency" and allo-centric predictions giving rise to "judgements of agency." Those two components are weighted according to their reliability and combined, generating a higher-level "sense of agency." We suggest that in schizophrenia a failure of corollary discharges to suppress self-generated inputs results in the absence of a "feeling of agency" and in a compensatory enhancement of allo-centric priors, which might underlie hallucinations, delusions of control but also, under certain circumstances, the enhancement of "judgments of agency." We discuss the consequences of such a model, and potential courses of action that could lead to its falsification.

Keywords: predictive processing, delusions, hallucinations, corollary discharge, psychosis, embodiment, agency

OPEN ACCESS

Edited by:
Thomas Parr,
University College London,
United Kingdom

Reviewed by:
Jun Tani,
Okinawa Institute of Science and Technology Graduate University, Japan
Michael Moutoussis,
University College London,
United Kingdom

****Correspondence:***
Philip R. Corlett
philip.corlett@yale.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *06 April 2020*
Accepted: *14 July 2020*
Published: *14 August 2020*

Citation:
Leptourgos P and Corlett PR (2020) Embodied Predictions, Agency, and Psychosis. Front. Big Data 3:27. doi: 10.3389/fdata.2020.00027

INTRODUCTION

In this article we will outline a computational account of perception and its disruption in psychosis. We will focus on the impact that actions have on the dynamics of perception. We will pay particular attention to how those dynamics may serve as grounds to infer agency over outcomes, and ownership of the body mediating the actions. Both ownership and agency are perturbed in people

with psychotic illnesses like schizophrenia. Such perturbations manifest as profound departures from the consensual sense of how bodies work, how intentions become manifest and how agency is ascribed. For example, someone with psychosis may believe that another agent is controlling their thoughts or actions against their will (passivity phenomena) and they may perceive agents alien to themselves talking inside their head [auditory verbal hallucinations (AVH)].

The framework we develop is grounded in notions of Bayesian inference and belief updating. Put simply, perceptions (of the self, the world, and their interaction) are inferences to the best explanation (abductions) of what would need to be the case in order for the data (from the world, body, and brain) to make sense. Those inferences are based on a model of what typically happens, combined with new data. We will argue that these inferences across sources of information (external world and internal milieu) are weighted by the reliability of those sources, if one stream becomes noisier, the others are given priority, and, given priority, beliefs about those sources can be self-reinforced and become rigidly immune to updating in light of new circumstances, just as we observe in the clinic from people with psychotic illnesses.

ROBOTS AND PREDICTIONS

We begin with a brief historical tour of the development of computational ideas relevant to action, perception, agency, and their disturbance in psychosis. Artificial intelligence—the construction and programming of intelligent machines, in cognitive science for the purpose of model building, theory construction, and hypothesis testing—has long been linked with psychiatry. In 1966, the computer scientist Joseph Weizenbaum created an early chatbot that searched for keywords in conversations conducted with human typers; if the human used one of those words, the program would use it in its reply. If not, it would offer a generic response. It was meant to mimic a psychotherapist (Weizenbaum, 1976). He named it ELIZA. In 1972, Kenneth Colby, then at Stanford created another program, PARRY—a bot that tried to model the behavior of a person with paranoia. That is, PARRY was constructed to behave as though espousing false beliefs of being harassed, subjugated, and persecuted, accused, mistreated, wronged, tormented, disparaged, vilified, and so on, by malevolent others, either specific individuals or groups. At the time, psychodynamic theories of paranoia prevailed—people were paranoid in order to protect themselves from the distress of shame and humiliation. Blaming others—the theory went—repudiated one's belief that they were to blame for an inadequacy. Parry has an interpretation module and an action module, and, through cycles of interaction with an interviewer, he progressively increments the weighting on beliefs that the interviewer has a poor opinion of him (Colby et al., 1971). Eliza and Parry interacted from different coasts of the US, via the nascent internet, and the results were amusing (Garber, 2014). Whilst they appeared to espouse knowledge and beliefs, these agents were really interacting via stimulus-response rules. They have only a shallow concept of "self," and many of the apparently paranoid behaviors that Parry evinced were hard coded based on actual patient responses. Parry and Eliza were far from having world knowledge, let alone knowledge of themselves as agents whose communicative acts impacted others.

More recently, the late Ralph Hoffman, who pioneered computational psychiatry, built and experimented with computational patients, network-based models of verbal cognition tasked with remembering brief narratives (Hoffman et al., 2011). Central to this function in the network, as in neural network models, is prediction error, the mismatch between input and retrieval. When model prediction errors were artificially elevated, the model misremembered narratives, inserting itself into stories. A perturbation of narrative agency. This approach was formally embodied by Yamashita and Tani (2012) who inserted a predictive coding architecture into a humanoid robot; with arms and a head [(Yamashita and Tani, 2012); see also Ohata and Tani, 2020, for a similar account of multimodal, imitative interaction of agents]. It had proprioceptive inputs from its arm joints and visual inputs that were modulated by the position vectors of its neck joints. The robot was confronted with a goal-object to be manipulated. Its task was to pick up and put down the object if it is in one position, and not if the object is in another position. Sometimes the experimenter would move the object between positions. The recurrent neural network that learned and executed the task was hierarchical and imbued with top-down predictions (intentions) and perceptual inputs. Mismatches between the intended and experienced events—prediction errors—were used to learn the task contingencies. If errant prediction errors were introduced to the network, the robot began to behave erratically, switching actions and perseverating-much like people with psychosis when making decisions under uncertainty. More recently, the same authors found that aberrant prediction errors can induce excessively strong priors in the same preparation (Idei et al., 2018). It is intriguing how, despite the mercurial increase in AI research, robotics has not tended to follow (Dennett, 1994). We posit that Tani et al. work does speak to embodiment, but perhaps not to the sense of conscious agency. We do not claim that a body is required for consciousness, since people with tetraplegia retain conscious experiences. They do however, experience agency differently, interacting with the world through effectors over which they retain some agency, like their eyes or mouths. This leads to an experience of dissociation and a much denuded sense of agency (Leggenhager et al., 2012). In psychosis, the agency change is different, it is a sense of too-little agency for some events (thoughts and actions) and too-much agency for others (outcomes, external events) (Moore and Fletcher, 2012). A kernel of the present paper is how strong priors and aberrant prediction errors can co-exist in the same brain and how those computational departures give rise to perturbed sense of agency over thoughts and actions and ultimately, hallucinations and delusions.

CONFLICTING ACCOUNTS

One influential theory of psychotic symptoms, hallucinations and delusions, posits that they are verbal thoughts, subvocal speech (in the case of hallucinations) or movements (in the case of passivity delusions) that are misattributed to an outside source (another agent that is communicating or controlling) (Jones and Fernyhough, 2007). This arises from compromised efference copy signals—"copies" of motor signals that are sent to sensory processing regions, rather than being sent to effectors, depositing a prediction of the expected sensory consequences of the action. Such self-induced stimulation is attenuated and may also underwrite agency attribution: I infer that I am the author of actions that proceed as expected, however, sufficient deviation from the predicted sensory consequences of actions invites the inference that another agent was involved. For hallucinations, there is some evidence for impaired efference copies of speech relating to hallucination severity, although by no means consistently. For passivity delusions, there is evidence for a failure in predictive motor cancellation that correlates with both hallucinations and delusions, in the realm of eye-movements and force-matching. If one conceives of efference copies as a kind of prior, these would be evidence for weak priors in people with psychosis that are related (perhaps) to the genesis of symptoms.

There is an alternative, based on the phenomenology of these symptoms. In particular their imperviousness to intersubjective data. That is, hallucinations and delusions do not respond well to the corrective influences of others. They are sustained despite overwhelming contradictory evidence. One might conceive of them not as relating to weak priors, but rather strong priors. If perception is an inferential process (Von Helmholtz, 1878), that inference that is optimized by prior knowledge about probable candidates (Von Helmholtz, 1866). The weighting of priors and current data is achieved by comparing their relative precision or inverse variance. If we are more confident in the data, they override our priors, if priors are more precise than sensory inputs, they will dominate inference and prediction errors will be ignored (Friston and Stephan, 2007; Friston and Kiebel, 2009; Feldman and Friston, 2010; Teufel et al., 2013). Hallucinations might arise when prior predictions exert an inordinate influence over perceptual inferences, creating percepts with no corresponding stimuli at all (Friston, 2005; Powers et al., 2016).

Indeed, in healthy volunteers who have undergone a training period that establishes an association between two stimuli, perceptual experiences of one stimulus (i.e., a tone) can occur in the absence of sensory input, conditional on the presentation of another stimulus (i.e., a visual stimulus) (Seashore, 1895), akin to a conditioned reflex (Pavlov, 1928; Ellson, 1941). More recently, visual-auditory conditioning has been employed to demonstrate that voice-hearing patients are significantly more susceptible to this effect than patients without hallucinations and controls (Kot and Serper, 2002). We recently showed that this effect is mediated by strong prior beliefs, that those priors are stronger in people who hallucinate, and that people with a diagnosed psychotic illness are less likely to update those prior beliefs in light of new evidence (Powers et al., 2017). Critically, the neural circuit underlying these conditioned phenomena—including superior temporal gyrus and insula—largely overlapped with the circuit engaged when patients report hearing voices in the scanner (Jardri et al., 2011; Powers et al., 2017). These studies underline the role of learning and, more specifically, a bias toward learned top-down information in the genesis of AVHs. Other studies, that probed the effect of high-level priors on bistable visual perception, came to similar conclusions (Schmack et al., 2013, 2017). Further support for this so-called strong prior account of hallucinations comes from findings that prior knowledge of a visual scene confers an advantage in recognizing a degraded version of that image (Teufel et al., 2015) and that patients at risk for psychosis—and, by extension, voice-hearing—were particularly susceptible to this advantage, and its magnitude correlated with hallucination-like percepts. Similarly, there is a version of this effect in audition; voice-hearing participants appear to have an enhanced prior for speech in degraded auditory stimuli even when not explicitly instructed (Alderson-Day et al., 2017). That is, speech is perhaps the most salient biological signal for our species, the auditory system of hallucination prone individuals may be pre-disposed to inferring speech. Likewise, the feeling of a lack of agency for our actions coupled with the experience that we are moving demands an explanation. All actions have a cause (internal or external) and agency typically accompanies self-generated movements. When agency is absent (i.e., the self is not the cause of the action), who or what might be causing that movement?

THE SENSE OF AGENCY

We constantly act to change our environment. Some actions are self-initiated, driven by our intentions and our expectations, while others are driven by external forces. For most of us, the distinction between voluntary and involuntary actions happens automatically, and is intimately related to the presence (or not) of a sense of agency. We define the sense of agency (SoA) as the experience of being in control of one's own actions and, through them, of events in the external world (Gallagher, 2000; Haggard, 2017). It constitutes, together with the sense of ownership [the experience that "my body" belongs to me (Tsakiris, 2017)], a key feature of self-consciousness (Braun et al., 2018) and underpins important concepts that define the human condition, such as free will and criminal responsibility (Haggard, 2017).

Despite its apparent unity, SoA consists of several components. An important distinction needs to be drawn between a *"feeling of agency"* (FoA) and a *"judgment of agency"* (JoA) (Synofzik et al., 2008a; Moore, 2016). The former can be experienced pre-reflectively and represents the non-conceptual feeling of control that colors our voluntary actions. On the contrary, JoA corresponds to a higher-level, conceptual construct that can be defined as "the ability to refer to oneself as the author of one's actions" (De Vignemont and Fourneret, 2004). The two levels of agency depend on each other [for example, FoA is a strong cue suggesting authorship of an action; "FoA is necessary but not sufficient for JoA" (Haggard and Tsakiris, 2009)] but, as recent studies have shown, they remain largely dissociable (Ebert

and Wegner, 2010; Strother et al., 2010; Dewey and Knoblich, 2014; Borhani et al., 2017).

Several implicit and explicit measures have been used to measure SoA, probing its different components. Implicit measures are considered as more appropriate for quantifying FoA, since they approach agency indirectly and avoid conscious judgments. One of the first implicit measures that was employed is *sensory attenuation*: the perceived intensity of sensations resulting from voluntary actions is diminished compared to sensations caused by involuntary (or external) actions [e.g., we cannot tickle ourselves (Blakemore et al., 2000b)]. Another implicit measure, considered as the hallmark of volition, is *intentional binding*: actions and the ensued outcomes are perceived closer together when the action is voluntary, resulting in a subjective contraction of time (Moore and Obhi, 2012). Indeed, in a series of studies Haggard et al. found that intentional binding occurred only in the case of voluntary actions, while involuntary actions evoked by Transcranial Magnetic Stimulation (TMS) of the motor cortex had the opposite result [repulsion (Haggard et al., 2002; Haggard and Clark, 2003)]. Since then, scientists have discovered links between binding and predictability (Moore and Haggard, 2008; Wolpe et al., 2013), associative learning [binding is enhanced by surprise (Moore et al., 2011a)], instrumental control (Borhani et al., 2017) and the fluency of action selection (Chambon et al., 2014).

On the other hand, explicit measures directly ask participants to make judgments about their agentic experience (Moore, 2016). In one example, participants are asked to make a hand movement and then see the same movement or a similar movement performed by another hand on a screen. In some cases and unbeknownst to the participant, a spatial or temporal distortion is added to the visual feedback. When asked whose hand they see on the screen, many participants misperceive the other hand as their own (especially in cases of no or small distortions), indicating the existence of a self-attribution bias (Farrer et al., 2003; Tsakiris et al., 2005; Hauser et al., 2011a). Other researchers asked the participants to make judgements about the feeling itself (Sidarus et al., 2013; Chambon et al., 2015). They found that parameters such as compatibility of priming, predictability and action-outcome delay profoundly affected participants' responses.

Disturbances of SoA are a common feature of psychotic disorders, such as schizophrenia (for a summary of the main empirical findings, see **Table 1**). Patients feel having no control over their actions and thoughts, which are instead controlled by external agents [passivity symptoms (Waters and Badcock, 2010)]. The presence of those passivity symptoms speaks to a diminished SoA in schizophrenia patients. However, carefully designed experiments found enhanced intentional binding (Haggard et al., 2003; Voss et al., 2010) and a stronger self-attribution bias in patients with schizophrenia (and passivity symptoms in particular) (Daprati et al., 1997; Franck et al., 2001), implying an exaggerated self-consciousness rather than a diminished sense of self (Hur et al., 2014). The apparent paradox is still not fully resolved, but evidence suggests a two-level impairment, namely an impairment in predictive components of agency (components related to processing occurring prior to action initiation (e.g., motor predictions, fluency of action selection etc.); possibly related to passivity symptoms), followed by an enhancement of retrospective processing (it includes processing that takes place after the action has been completed and feedback has been received; perhaps resulting in over-attribution) (Synofzik et al., 2010; Voss et al., 2010).

The computational underpinnings of agency have also been lively debated over the past 30 years. According to the influential *comparator model* (Feinberg, 1978; Blakemore et al., 2000a; Blakemore and Frith, 2003), SoA relies on the motor system that is responsible for initiating and controlling self-generated movements, based on the principles of optimal control theory (Wolpert et al., 1995; Wolpert and Ghahramani, 2000). More particularly, the brain predicts the sensory consequences of self-initiated actions through the use of *forward models* (Wolpert and Kawato, 1998). A copy of the motor prediction [*corollary discharge*; often called *efference copy* (Feinberg, 1978)] is sent to the sensory areas, suppressing predictable inputs (proprioceptive but also visual, auditory etc.). This sensory attenuation of self-generated inputs (discussed above) ultimately gives rise to the feeling that one is in control of their own actions.

Despite its success, several criticisms against the comparator model have been raised, largely based on its inability to account for JoA (e.g., Synofzik et al., 2008a). According to the theory of *apparent mental causation*, put forward by Wegner and Wheatley, SoA does not rely on the motor signals that initiated the action, but on generic inferential processes (Wegner and Wheatley, 1999). In a nutshell, this theory suggests that (1) if an action is preceded by an intention, (2) if the action is compatible with that intention and (3). if the intention is the most likely cause of the action, then the action is attributed to one's self. Intriguingly, this theory is not based on "private" mechanisms (such as the motor signals) and thus, it can be generalized to other peoples' actions. More recently, several theorists tried to combine the above models, bridging the gap between motor and inferential processes and, more generally, between FoA and JoA (Synofzik et al., 2008a,b; Moore and Fletcher, 2012; Moutoussis et al., 2014; Kahl and Kopp, 2018; Legaspi and Toyoizumi, 2019).

RECONCILING CONFLICTING ACCOUNTS

In the previous sections we argued that two different riddles have been puzzling researchers for decades. Namely:

- Are hallucinations and delusions due to strong or weak priors [i.e., strong priors (Sterzer et al., 2018; Corlett et al., 2019) vs. weak corollary discharge (Blakemore et al., 2002; Thakkar et al., 2017) and a loss of agency for one's inner speech (Jones and Fernyhough, 2007)]?
- Relatedly, do schizophrenia patients have an exaggerated or a diminished SoA?

Paradoxically, in both cases there is evidence supporting strong and weak priors, weak corollary discharge, misattributed inner speech, exaggerated, and diminished agency (though typically not at the same time in the same people with psychosis).

TABLE 1 | Sense of agency/ownership (implicit and explicit measures) in psychosis: main empirical findings.

References	Population (sample size)	Paradigm	Main findings
Malenka et al. (1982)	SCZ (14)	Tracking task (Error corrections)	SCZ: Fewer error corrections without external (visual) cues
Frith and Done (1989)	SCZ + AP (23) (P+: 10; P–: 13)	Motor task (Error corrections)	SCZ: Fewer error corrections without external (visual) cues
Daprati et al. (1997)	SCZ (30) (H+: 13; DC+: 7; H+DC+: 6; H–DC–: 10)	Recognition task ("Is that my hand on the screen?")	H+, DC+: More false self-attributions
Blakemore et al. (2000a)	SCZ (23) + AD (18) (H+: 17; H+P+: 6; H–P–: 24)	Sensory attenuation task (tactile stimulation)	H+, P+: No sensory attenuation of self-produced tactile sensations
Franck et al. (2001)	SCZ (24) (P+: 6; P–: 18)	Recognition task ("Is that my hand on the screen?")	P+: More false self-attributions
Delevoye-Turrell et al. (2002)	SCZ (16) (DC+: 6; DC–: 10)	Force adjustment task	DC+, DC–: No improvement of efficiency of motor response in self- vs. externally- imposed condition
Haggard et al. (2003)	SCZ (8)	Intentional binding task	SCZ: Stronger binding between actions and outcomes
Allen et al. (2004)	SCZ (28) (H+D+: 15; H–D–: 13)	Recognition task ("Is this my voice?")	H+D+: More misidentifications of their own speech as alien—correlation with severity of hallucinations
Knoblich et al. (2004)	SCZ (27)	Motor task (implicit—explicit error corrections)	SCZ (with symptoms): Impaired explicit detection of action-outcome mismatches/intact implicit corrections
Lindner et al. (2005)	SZC (14)	Sensory attenuation task (Smooth-pursuit eye-movement task)	SCZ: Less sensory attenuation (stronger reafference) —correlation with severity of delusions of control
Shergill et al. (2005)	SCZ (19)	Sensory attenuation task (Force-matching task)	SCZ: Less sensory attenuation (less underestimation of self-generated force)
Synofzik et al. (2010)	SCZ (20)	Task 1: Detection of discrepancies between action (pointing) and distorted visual feedback Task 2: Estimation of direction of pointing with or without distorted visual feedback	SCZ: Task 1—Higher thresholds for detecting action-outcome discrepancies; Task 2—More adaptation of estimates to feedback; More variable estimates; Variability of estimates (in the absence of feedback) correlated with delusions of control and detection thresholds from task 1
Teufel et al. (2010)	CTR (30)	Sensory attenuation task (Force-matching task)	Participants with higher delusion-proneness (PDI score) exhibited weaker sensory attenuation
Voss et al. (2010)	SCZ (24)	Intentional binding task	SCZ: Impaired predictive component of action awareness (weaker effect of outcome predictability—correlated with positive symptoms) —enhanced retrospective component (presence of the outcome)
Hauser et al. (2011a)	SCZ (30); PP (30)	Recognition task ("Did I produce this tone?")	Both SCZ and PP: More false self-attributions—self-attribution bias correlated with passivity symptoms
Hauser et al. (2011b)	PP (30)	Intentional binding task	PP: Stronger intentional binding—both predictive and retrospective influences were stronger—predictive influences correlated with ego-psychopathology (IPP score)
Moore et al. (2011b)	CTR_Ket (14)	Intentional binding task	Ketamine enhances binding—correlation with aberrant bodily experiences
Thakkar et al. (2011)	SCZ (24)	Rubber Hand Illusion (RHI)	SCZ: Stronger RHI (both implicitly and explicitly measured)—self-reported strength of RHI correlated with schizotypy in CTR
Maeda et al. (2012)	SCZ (30)	Agency attribution task	SCZ: Excessive sense of agency (even when outcomes precede actions)
Renes et al. (2013)	SCZ (23)	Agency attribution task (explicit condition: intentions/implicit condition: priming)	SCZ: Enhanced self-agency in explicit condition (not different from CTR)—Less enhancement than CTR in implicit condition
Hur et al. (2014)	Meta-analysis—25 studies SCZ (690)		Self-disturbance in SCZ: distortions in body-ownership, self of agency (enhanced) and self-reported subjective experiences
Moore and Pope (2014)	CTR (35)	Agency attribution task with video stimuli	Presence of intentionality bias. The bias is stronger in individuals with stronger schizotypal traits
Koreki et al. (2015)	SCZ (30)	Agency attribution task	SCZ: Excessive sense of agency (even for action-outcome delays longer than 1s)

(Continued)

TABLE 1 | Continued

References	Population (sample size)	Paradigm	Main findings
Garbarini et al. (2016)	SCZ (20)	Bimanual coupling task (bimanual condition: participants draw lines with one hand and circles with the other—modified condition: participant draws lines with one hand while observing examiner drawing circles)	SCZ: Same interference effects in bimanual condition—stronger interference in modified condition
Lemaitre et al. (2016)	CTR (ST+: 27; ST–: 27)	Sensory attenuation task (tactile stimulation)	Self-applied tactile stimulations are felt to be more ticklish by healthy individuals high in schizotypal traits—self-tickling was associated with passivity experiences
Voss et al. (2017)	SCZ (14)	Agency attribution task + priming	SCZ: Similar effects of priming on motor performance—no effect of priming on sense of agency (contrary to CTR)
Whitford et al. (2017)	CTR (110)	Sensory attenuation task (tactile stimulation)	Participants with stronger schizotypal traits (SPQ score) exhibited weaker sensory attenuation
Graham-Schmidt et al. (2018)	SCZ (51) (Current P+: 20; Past P+: 10; P–: 21)	Projected Hand Illusion (PHI)	P+ (current or past): Less difference in agency between active and passive movements when assessing agency over their own hand

SCZ, Schizophrenia patients; AP, Affective Psychosis; AD, Affective disorder; PP, Prodromal Patients; CTR, Controls; CTR_Ket, Controls given Ketamine; H+, With hallucinations; H–, Without hallucinations; D+, With delusions; D–, Without delusions; DC+, With delusions of control; DC–, Without delusions of control; P+, With passivity symptoms; P–, Without passivity symptoms; FTD+, With formal thought disorder; ST+, High schizotypal traits; ST–, Low schizotypal traits.

In this section we advance a conceptual model which, we believe, can reconcile those (seemingly) contradictory accounts. We argue that there is more than one generative model in the brain, and that ego- and allo-centric models operate in tandem. In brief, there are inferences (related to actions) that need to represent and account for the impact of self on perception (ego-centric) and there are inferences that do not need such accounting (allo-centric). There may be a precision-weighted trade-off between which source is drawn upon for inference, especially in the case of agency attribution. Such a trade-off would allow for aberrant corollary discharges and strong priors in the same individual, both of which contribute to symptom genesis. Additionally, by postulating that each one of the two hierarchies is responsible for a different level of agency attribution, our model can predict both exaggerated and diminished SoA, depending on the experimental context. We note that a detailed mathematical description is beyond the scope of this paper and will be presented in future publications.

The Model

Our model is illustrated in **Figure 1**. It consists of two hierarchies operating in parallel: An ego-centric hierarchy, predicting self-generated inputs, and an allo-centric hierarchy, implementing more general inferences regarding the state of the world. Interestingly, the 2 hierarchical systems are related to each other, as they receive bottom-up information from the same sensory systems [e.g., retina and primary visual cortex in the case of visual inputs; proprioceptors and cerebellum in the case of proprioception (Shergill et al., 2014)].

The ego-centric system is part of a sensorimotor loop, that controls and optimizes the trajectories of movements (Wolpert and Ghahramani, 2000). Copies of the motor commands (i.e., efference copies) are transformed into motor predictions about the sensory consequences of self-generated actions through the use of internal predictors, the forward models (Wolpert and Kawato, 1998). Those motor predictions (i.e., corollary discharges) are then weighted according to their reliability (w_{ego}) and sent to sensory areas, where they attenuate precision-weighted (w_V, w_P etc.) self-generated inputs (primarily proprioceptive but also visual, auditory etc.; Wolpert et al., 1995; Blakemore and Frith, 2003; Körding and Wolpert, 2004). Importantly, the motor predictions and their precision can be modulated by various factors such as intentions or cues preceding action initiation (priming effects or fluency of action selection). For example, fluent selection of the appropriate action might have profound effects on the strength of the efferent signals (Chambon et al., 2014).

The allo-centric system on the other hand implements more generic predictive processing based on the principles of hierarchical Bayesian inference. Very briefly, that means that the allo-centric system learns and represents causal models of the world and inverts those models to estimate the most probable cause of the sensory input [self-produced or not (Von Helmholtz, 1866; Clark, 2013)]. According to predictive coding theory, high-level predictions (weighted according to their reliability w_{allo}) explain away sensory inputs (w_V,w_P etc.), in the same way motor predictions suppress self-generated inputs (Friston and Kiebel, 2009). When there is a mismatch between predictions and inputs, a prediction error signal is generated which updates the current model. Importantly, this constructive view of perception implies that percepts are not pure representations of sensory inputs, instead they are biased by prior knowledge, which might be learnt through experience [e.g., empirical priors (Friston, 2009)] or hard-coded through evolution [e.g., "light comes from above" (Mamassian and Landy, 1998; Dobbins and Grossmann,

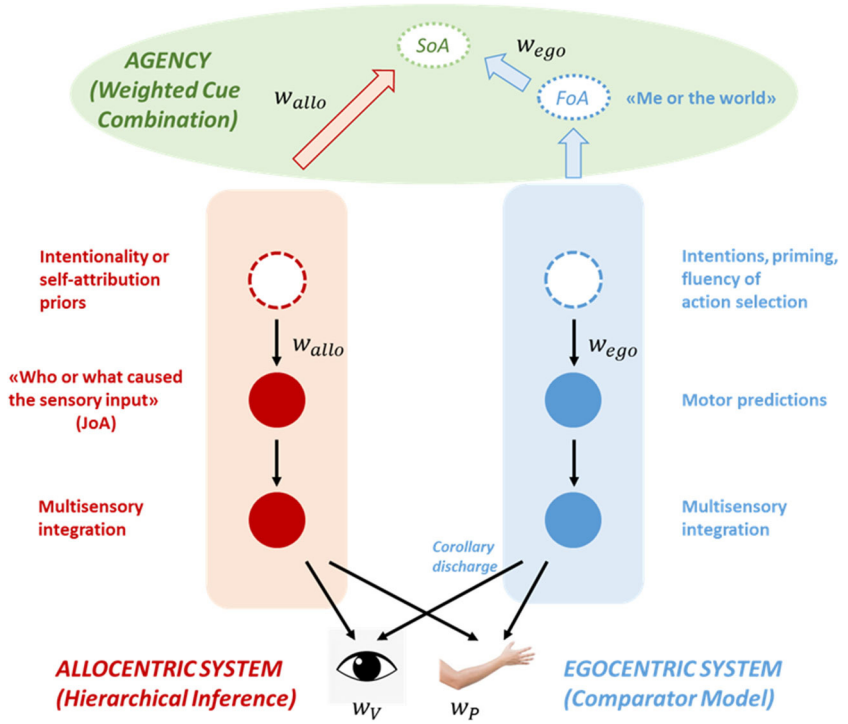


FIGURE 1 | An illustration of the model. The model consists of two hierarchies, an ego-centric and an allo-centric system, that operate in tandem and interact at the sensory level. The ego-centric system (in blue) is part of a sensorimotor loop and implements a comparator model. A copy of the motor command (transformed via a forward model into a motor prediction about the sensory outcomes of the action) is sent to the sensory areas where it suppresses self-generated (i.e., predictable) inputs. Motor predictions can be modulated by higher-level factors such as intentions or the fluency of action selection. The allo-centric system (in red) represents generative causal models of the world, including the self among the potential causes. According to predictive coding, allo-centric predictions (like motor predictions) explain away predictable inputs, but unlike the ego-centric system those inputs are not necessarily self-generated. Allo-centric predictions are also modulated by higher-level priors such as an intentionality or a self-attribution bias. Both types of predictions and the sensory inputs are weighted according to their reliability (w_{ego}, w_{allo}, and w_V, w_P, respectively). Crucially, both systems make inferences about different levels of agency. The ego-centric system implements a private mechanism that makes a self-world distinction and gives rise to a feeling of agency (FoA) when motor predictions and inputs are in good match. The allo-centric system on the other hand generates judgments of agency (JoA) based on generic inferential mechanisms, by comparing multiple hypothesis about the cause of a certain outcome ("Me" vs. "External agent" vs. "External non-agentic cause" vs…). The different components of agency are then fed-forward to an agency-attribution system (in green), where they are combined according to a weighted cue combination mechanism that gives rise to a higher-level sense of agency (SoA).

2010)]. It's worth noting that learning can be driven both by the reliability of the cues and by uncertainty (Corlett, 2020).

Crucially, both systems contain the necessary machinery to make inferences about the contribution of the self in the generation of the inputs and thus, about agency (Wegner and Wheatley, 1999; Blakemore and Frith, 2003). The ego-centric system is an implementation of Frith's comparator model (Blakemore et al., 2000b). More particularly, ego-centric (motor) predictions suppress sensory inputs only in case they are predictable, that is, if they are self-generated. Consequently, sensory attenuation should be followed by a feeling that one is control of their own actions, in other words, they should experience a FoA. We should highlight here that this is a "private" mechanism, only applicable to one's self (Synofzik et al., 2008a; Carruthers, 2009). That means that the ego-centric system does not have the necessary mechanisms to attribute agency to someone else; it can only decide between "me" and "the world."

The allo-centric system relies on more eight general inferential mechanisms, therefore it can choose between different internal and external causes ("me," "you," an object etc.), potentially underwriting judgments of agency (JoA). Those inferences can rely on sensory inputs (e.g., movement of a hand, moving lips etc.) but also on priors regarding the intentions of others. Furthermore, those agency-related inferences might also be driven by hardwired biases such as the intentionality bias (Rosset, 2008; Sidarus et al., 2013) or a self-attribution bias (Farrer et al., 2003; Tsakiris et al., 2005; Hauser et al., 2011a). In the context of Bayesian inference, those biases can be conceptualized as additional priors or hyperpriors (priors on hyperparameters that control the shape of the prior distributions). Although those additional priors can reduce the accuracy of the agency attribution mechanism, they might enhance social bonding, underpin "theory of mind" or increase self-esteem (see also Garety and Freeman, 1999; Schwarz et al., 2016).

Following previous theoretical suggestions (Synofzik et al., 2008a; Moore and Fletcher, 2012; Kahl and Kopp, 2018), we

postulate that FoA and JoA are combined to generate a higher-level SoA via a precision-weighted cue combination mechanism, where the 2 weights can be related to the precision of the ego-centric and allo-centric predictions, respectively. For example, a partial attenuation of the input by the ego-centric predictions might result in a lack of a FoA, which can in turn override the allo-centric intentionality priors (and a potentially positive JoA), resulting in the belief that we are not the author of the action.

In the next section, we describe the implications of the model in the case of schizophrenia. In particular, we suggest that the interactions between and within the two hierarchies of inference can reconcile the apparent contradictions (Sterzer et al., 2018; Corlett et al., 2019).

Schizophrenia[1]: From Weak Motor Predictions to Strong Allo-Centric Predictions

There is an abundance of evidence that interactions between motor and perceptual systems are crucial for both functions (Faivre et al., 2015). A well-functioning perceptual system (i.e., a system that attributes precise weights to priors and sensory inputs, according to their reliability) makes accurate perceptual decisions, which in turn can lead to meticulous adjustments of the self-generated movements, through the operation of sensorimotor loops. Vice versa, intact corollary discharges explain away the unnecessary self-induced sensory signals, preventing them from affecting allo-centric inferences (**Figure 2A**). An interesting example of this fine-tuned interaction is saccadic suppression and the ensued visual stability during eye-movements (Melcher, 2011): although we make several saccades every second, whose peak speed can reach several hundreds of degrees/sec, we perceive no changes in our visual field, an effect that is usually attributed to efferent inhibitory motor signals (corollary discharge) (Cavanaugh et al., 2016). Importantly, the optimal integration of the allo- and ego-centric predictions also results in precise agency-estimates, based on the accurate calculation and combination of the FoA and JoA.

What happens if we selectively impair corollary discharge signals, as described in schizophrenia (Blakemore et al., 2002; Synofzik et al., 2010; Thakkar et al., 2017)? Motor predictions cannot explain away self-generated signals, resulting in a reduced sensory attenuation of those sensations (**Figure 2B**; Blakemore et al., 2000a; Shergill et al., 2005) and a diminished FoA. That explains why patients with schizophrenia do not feel in control of their own actions, however it does not explain why they attribute their actions to an external agent (Frith, 2005).

Ego-centric and allo-centric hierarchies work in tandem. We argue that impairments in one system (e.g., weak corollary discharge) have a profound effect in the opposite system as well (Corlett et al., 2019; Thakkar and Rolfs, 2019). In the case of schizophrenia patients, the un-attenuated self-generated sensory signals would penetrate in the allo-centric hierarchy, flooding it with noisy, inherently unpredictable information (e.g., rapidly changing visual inputs during saccadic movements; see also Seal et al., 2004; Jones and Fernyhough, 2007; Alderson-Day and Fernyhough, 2015) and also resulting in low level perceptual abnormalities (e.g., blurred images, changes in perception of size or color etc.). Various experimental findings corroborate this idea: first, patients exhibit deficient saccadic suppression, which results in unstable visual images during movement (pseudo-movements) (Krekelberg, 2010; Thakkar and Rolfs, 2019); second, self-generated, subvocal speech, picked by throat microphones, has been causally associated with certain types of AVH (Gould, 1950; Bick and Kinsbourne, 1987), suggesting that self-generated stimuli receive special attention and are mis-processed by patients; third, when people report AVH in the scanner, their speech network (including both speech production and reception areas) is engaged (Jardri et al., 2011). This penetration gives rise to strong prediction error signals, which are propagated toward higher levels, constantly updating the internal models. Additionally, given the tight connection between saccades and spatial attention, impaired corollary discharge signals might also give rise to attentional problems, including the aberrant salience attributed to random stimuli in the environment (Thakkar and Rolfs, 2019). In both cases, the world would seem unstable, unpredictable and strange. We suggest that the allo-centric system compensates for the overwhelming bottom-up signals by increasing the precision of high-level allo-centric priors (Adams et al., 2013; Schmack et al., 2013, 2017; Powers et al., 2017; Sterzer et al., 2018; Corlett et al., 2019). This compensatory mechanism would alleviate the strong impact of the self-generated signals by increasing the relative contribution of the priors in allo-centric inferences, resulting in more stable and less chaotic percepts. Despite its beneficial effect, this overreliance on priors also renders the system more vulnerable to hallucinations (**Figure 2B**). Indeed, auditory hallucinations are one of the most prominent symptoms in schizophrenia and have been repeatedly associated with strong priors (Teufel et al., 2015; Powers et al., 2016, 2017). Can strong priors also explain the content of hallucinations and delusions (e.g., predominantly negative content of AVH, technical delusions etc.)? This is not an unreasonable speculation, especially if also take into account the affective and cultural forces that "shape" those priors (Škodlar et al., 2008; Laroi et al., 2014).

This enhancement of allo-centric priors also has significant effects on the SoA. Combined with the down-regulation of the motor predictions, it means that the JoA gains a particular significance, compared to FoA. But JoA is subject to various biases, including an intentionality bias (Rosset, 2008). This means that individuals that overweight priors would have a stronger tendency to attribute actions to hidden intentions, thus perceive volitional behaviors even when there are none. Taken together, they explain the phenomenology of delusions of control, where

[1]Schizophrenia is a heterogeneous disorder, characterized by positive (psychotic), negative and cognitive symptoms. The model outlined in this paper is an effort to understand mechanistically the positive symptoms (more particularly, hallucinations and delusions of control) and not schizophrenia as a whole. Other common symptoms of schizophrenia, including negative symptoms and other types of delusions (e.g. persecutory delusions), are likely to be underwritten by different mechanisms.

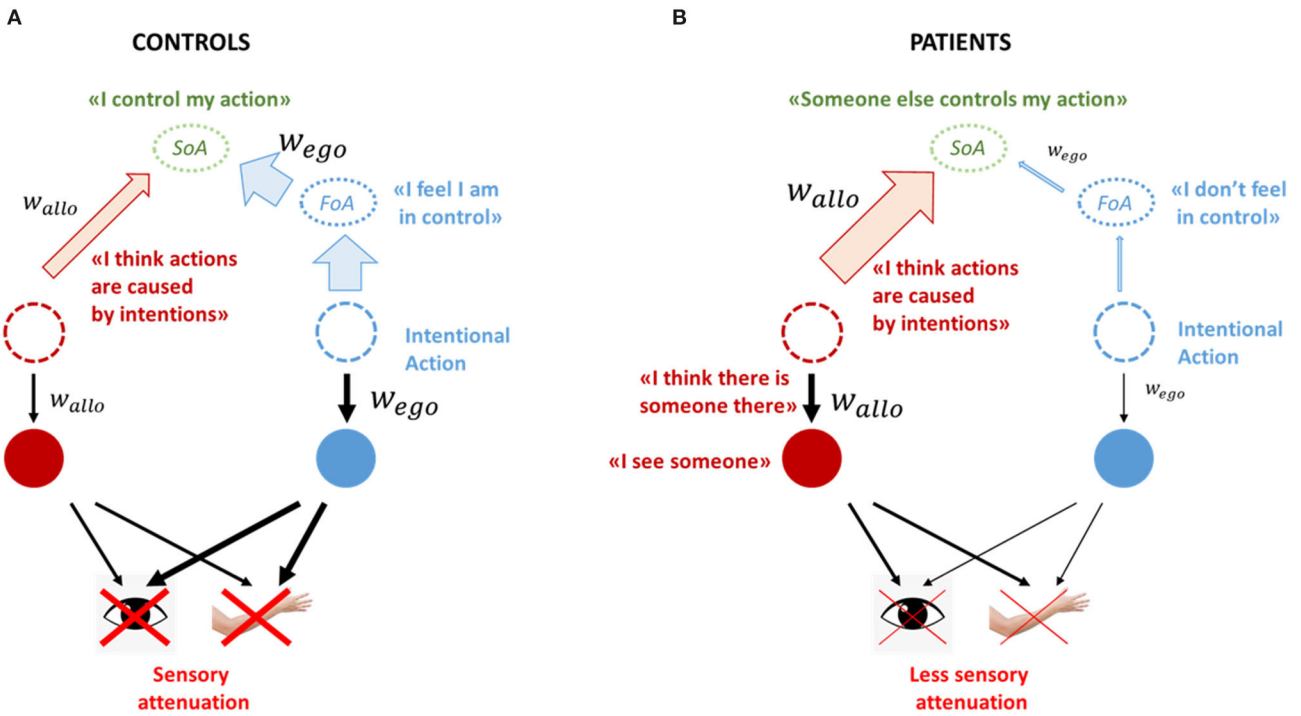


FIGURE 2 | Healthy controls vs. Schizophrenia patients. **(A)** In a well-functioning system, predictions (both ego-centric and allo-centric) are weighted according to their reliability. When an intentional action is initiated, strong corollary discharge signals explain-away self-generated inputs, which in turn give rise to a FoA and a sense that one is in control of their actions. Within the allo-centric system, that results in optimal perceptual inferences **(B)** When motor predictions are under-weighted (e.g., in schizophrenia), self-generated inputs cannot be explained away, resulting in a feeling that one is not in control of their actions. The unsuppressed inputs flood into the allo-centric system, which is overflowed with noisy and inherently unpredictable information. To compensate for that, it increases the weight of high level allo-centric priors, including agency-related priors such as the intentionality bias. Strong priors have an effect both on perceptual decision making and on agency-attribution: first, percepts are mainly driven by priors, rendering the system susceptible to hallucinations. Second, the enhanced intentionality bias, combined with the lack of a FoA, bring about the false belief that an external agent is in control of one's own actions, i.e., a delusion of control.

people do not feel in control of their own actions and attribute them to external forces (Frith, 2005).

Interestingly, the same impairments can also explain the opposite pattern, notably the tendency of schizophrenia patients with passivity symptoms to over-attribute certain actions to themselves in recognition tasks (Daprati et al., 1997; Franck et al., 2001). The key observation here is that in those tasks ego-centric predictions are largely irrelevant; a FoA is dissociated from the perceptual decision "is this my hand." In this case, a SoA (and consequently the perceptual decision) depends first and foremost on allo-centric JoA. But JoA is also subject to a self-attribution bias [(Garety and Freeman, 1999); the intentionality bias is also at play], which is enhanced due to the overweighted priors. Consequently, patients can over-attribute and under-attribute actions to themselves, depending on the experimental context. Similar arguments can be put forward to explain delusions of reference (Maeda et al., 2012), while it's an open question whether similar mechanisms could explain other first-rank symptoms such as thought insertion or made feelings (Vosgerau and Newen, 2007; Frith, 2012).

In short, we described a conceptual model that reconciles contradictory accounts of schizophrenia, namely whether patients over-weight or under-weight their priors, and whether they have an exaggerated or a diminished SoA. The model can explain various state symptoms (symptoms that manifest themselves during full-blown psychotic episodes, such as hallucinations, delusions of control or even low-level perceptual abnormalities), it remains unclear though whether similar mechanisms could also explain trait symptoms [more permanent features of schizophrenia, also found in first-degree relatives and high-risk populations (Adams et al., 2013)] and, more importantly, different phases of the disorder, such as the prodromal phase. In the next section we describe some further predictions of the model.

EXPLANATORY POWER AND NOVEL PREDICTIONS

The combined impaired-corollary discharge and strong-priors account that we outlined above makes some additional predictions, some of them novel, meaning that it is a highly falsifiable theory. That said, given the conceptual nature of the described model, our predictions should be made with caution.

First, it is compatible with data suggesting both compromised motor predictions (Lindner et al., 2005; Synofzik et al.,

2010; Thakkar et al., 2017) and overly strong priors (Powers et al., 2017). Importantly, because of the assumed causal link between the two, we expect an anti-correlation within the same individuals (Corlett et al., 2019); e.g., participants with less sensory attenuation and stronger re-afferent signals should also report more conditioned hallucinations. Stronger evidence in favor of our theory could be obtained from causal, virtual lesion studies such as TMS studies: stimulation of regions critically involved in ego-centric inferences such as cerebellum (Blakemore et al., 1998; Synofzik et al., 2008c) or the temporo-parietal junction (TPJ) (Hughes, 2018) should engender hallucinations in participants (Arzy et al., 2006).

More generally, our theory suggests that failures of the ego-centric system would render the perceptual system more susceptible to false percepts and hallucinations. Interestingly, recent work suggests that sensorimotor conflicts induced by a robotic system decrease the capacity to adapt confidence to task performance (metacognitive failure), increase intentional binding (potentially due to an enhanced JoA) (Faivre et al., 2020) and generate a feeling of presence (Blanke et al., 2014).

Finally, our theory makes several predictions regarding SoA and its impairments in schizophrenia and in related psychotic disorders (Hauser et al., 2011a,b; Moore et al., 2011b). Primarily, our theory predicts an anti-correlation between FoA and JoA (and the related explicit or implicit measures of FoA and JoA) within the same participants. For example, one might expect decreased sensory attenuation (an implicit measure of FoA; Shergill et al., 2005; Teufel et al., 2010) to correlate with increased self-over-attribution in recognition tasks (Daprati et al., 1997; Franck et al., 2001). Intriguingly, one might also expect judgments of ownership (JoO), whose cognitive and computational mechanisms partly overlap with those of JoA (Tsakiris, 2017), also to anti-correlate with FoA. For example, vulnerability to the rubber hand illusion [(Tsakiris and Haggard, 2005); an increased vulnerability of the RHI has been observed in schizophrenia patients (Thakkar et al., 2011)] should correlate with less sensory attenuation. Ultimately, the present theory also explains several observations about intentional binding, such as the reduced effect of priming (Voss et al., 2017) and the enhanced effect of retrospective processing (Voss et al., 2010) in schizophrenia patients, while it also predicts a decreased effect of the fluency of action selection in the same populations (Chambon et al., 2014).

SUMMARY AND CONCLUSIONS

This paper outlines an account of inference and agency that reconciles several conflicting lines of evidence. Ego-centric and allo-centric models operate in tandem, making up the machinery required for attaining self-other distinction and thus, SoA. Ego-centric models implement corollary discharge signals that cancel out the effects of self-generated actions, subserving FoA. Allo-centric models compare several hypothesis regarding the causes of sensory inputs (including the self among the potential causes), giving rise to JoA. The different levels of agency are weighted according to their reliability and combined, ultimately forming a higher-level SoA. In schizophrenia, a failure of corollary discharges to suppress self-generated inputs results in the absence of a FoA and in a (compensatory) enhancement of allo-centric priors, which might underlie hallucinations, delusions of control but also, under certain circumstances, the enhancement of JoA.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

PL and PC conceived the model and wrote the paper. All authors contributed to the article and approved the submited version.

FUNDING

This work was supported by the Yale University Department of Psychiatry, the Connecticut Mental Health Center (CMHC) and Connecticut State Department of Mental Health and Addiction Services (DMHAS). It was funded by an IMHRO/Janssen Rising Star Translational Research Award and NIMH R01MH12887. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

REFERENCES

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., and Friston, K. J. (2013). The computational anatomy of psychosis. *Front. Psychiatry* 4:47. doi: 10.3389/fpsyt.2013.00047

Alderson-Day, B., and Fernyhough, C. (2015). Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychol. Bull.* 141, 931–965. doi: 10.1037/bul0000021

Alderson-Day, B., Lima, C. F., Evans, S., Krishnan, S., Shanmugalingam, P., Fernyhough, C., et al. (2017). Distinct processing of ambiguous speech in people with non-clinical auditory verbal hallucinations. *Brain* 140, 2475–2489. doi: 10.1093/brain/awx206

Allen, P. P., Johns, L. C., Fu, C. H. Y., Broome, M. R., Vythelingum, G. N., and Mcguire, P. K. (2004). Misattribution of external speech in patients with hallucinations and delusions. *Schizophr. Res.* 69, 277–287. doi: 10.1016/j.schres.2003.09.008

Arzy, S., Seeck, M., Ortigue, S., Spinelli, L., and Blanke, O. (2006). Induction of an illusory shadow person. *Nature* 443:287. doi: 10.1038/443287a

Bick, P. A., and Kinsbourne, M. (1987). Auditory hallucinations and subvocal speech in schizophrenic patients. *Am. J. Psychiatry* 144, 222–225. doi: 10.1176/ajp.144.2.222

Blakemore, S.-J., and Frith, C. (2003). Self-awareness and action. *Curr. Opin. Neurobiol.* 13, 219–224. doi: 10.1016/S0959-4388(03)00043-6

Blakemore, S.-J., Wolpert, C. A. D., and Frith, C. (2000b). Why can't you tickle yourself? *Neuroreport* 11, 11–16. doi: 10.1097/00001756-200008030-00002

Blakemore, S.-J., Wolpert, D. M., and Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nat. Neurosci.* 1, 635–640. doi: 10.1038/2870

Blakemore, S. J., Smith, J., Steel, R., Johnstone, E., and Frith, C. (2000a). The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: evidence for a breakdown in self-monitoring. *Psychol. Med*. 30, 1131–1139. doi: 10.1017/S0033291799002676

Blakemore, S. J., Wolpert, D., and Frith, C. (2002). Abnormalities in the awareness of action. *Trends Cogn. Sci.* 6, 237–242. doi: 10.1016/S1364-6613(02)01907-1

Blanke, O., Pozeg, P., Hara, M., Heydrich, L., Serino, A., Yamamoto, A., et al. (2014). Neurological and robot-controlled induction of an apparition. *Curr. Biol.* 24, 2681–2686. doi: 10.1016/j.cub.2014.09.049

Borhani, K., Beck, B., and Haggard, P. (2017). Choosing, doing, and controlling: implicit sense of agency over somatosensory events. *Psychol. Sci*. 28, 882–893. doi: 10.1177/0956797617697693

Braun, N., Debener, S., Spychala, N., Bongartz, E., Sörös, P., Müller, H. H. O., et al. (2018). The senses of agency and ownership: a review. *Front. Psychol*. 9:535. doi: 10.3389/fpsyg.2018.00535

Carruthers, G. (2009). Commentary on synofzik, vosgerau and newen 2008. *Conscious Cogn*. 18, 515–520. doi: 10.1016/j.concog.2008.05.006

Cavanaugh, J., Berman, R. A., Joiner, W. M., and Wurtz, R. H. (2016). Saccadic corollary discharge underlies stable visual perception. *J. Neurosci.* 36, 31–42. doi: 10.1523/JNEUROSCI.2054-15.2016

Chambon, V., Moore, J. W., and Haggard, P. (2015). TMS stimulation over the inferior parietal cortex disrupts prospective sense of agency. *Brain Struct. Funct*. 220, 3627–3639. doi: 10.1007/s00429-014-0878-6

Chambon, V., Sidarus, N., and Haggard, P. (2014). From action intentions to action effects: how does the sense of agency come about? *Front. Hum. Neurosci*. 8:320. doi: 10.3389/fnhum.2014.00320

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Colby, K. M., Weber, S., and Hilf, D. F. (1971). Artificial paranoia. *Artif. Intell.* 2, 1–25. doi: 10.1016/0004-3702(71)90002-6

Corlett, P. (2020). Predicting to perceive and learning when to learn. *Trends Cognitive Sci.* 24:259–260. doi: 10.1016/j.tics.2019.12.005

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., and Powers, A. R. (2019). Hallucinations and strong priors. *Trends Cogn. Sci.* 23, 114–127. doi: 10.1016/j.tics.2018.12.001

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J., et al. (1997). Looking for the agent: an investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognitiongnition* 65, 71–86. doi: 10.1016/S0010-0277(97)00039-5

De Vignemont, F., and Fourneret, P. (2004). The sense of agency: a philosophical and empirical review of the "Who" system. *Conscious. Cogn*. 13, 1–19. doi: 10.1016/S1053-8100(03)00022-9

Delevoye-Turrell, Y., Giersch, A., and Danion, J.-M. (2002). A deficit in the adjustment of grip force responses in schizophrenia. *Neuroreport* 13, 1537–1539. doi: 10.1097/00001756-200208270-00010

Dennett, D. (1994). The practical requirements for making a conscious robot. *Philos. Trans.* 349, 133–146. doi: 10.1098/rsta.1994.0118

Dewey, J. A., and Knoblich, G. (2014). Do implicit and explicit measures of the sense of agency measure the same thing? *PLoS ONE* 9:e110118. doi: 10.1371/journal.pone.0110118

Dobbins, A. C., and Grossmann, J. K. (2010). Asymmetries in perception of 3D orientation. *PLoS ONE* 5:e9553. doi: 10.1371/journal.pone.0009553

Ebert, J. P., and Wegner, D. M. (2010). Time warp: authorship shapes the perceived timing of actions and events. *Conscious. Cogn*. 19, 481–489. doi: 10.1016/j.concog.2009.10.002

Ellson, D. G. (1941). Hallucinations produced by sensory conditioning. *J. Exp. Psychol*. 28, 1–20. doi: 10.1037/h0054167

Faivre, N., Salomon, R., and Blanke, O. (2015). Visual consciousness and bodily self-consciousness. *Curr. Opin. Neurol*. 28, 23–28. doi: 10.1097/WCO.0000000000000160

Faivre, N., Vuillaume, L., Bernasconi, F., Salomon, R., Blanke, O., and Cleeremans, A. (2020). Sensorimotor conflicts alter metacognitive and action monitoring. *Cortex* 124, 224–234. doi: 10.1016/j.cortex.2019.12.001

Farrer, C., Franck, N., Paillard, J., and Jeannerod, M. (2003). The role of proprioception in action recognition. *Conscious. Cogn*. 12, 609–619. doi: 10.1016/S1053-8100(03)00047-3

Feinberg, I. (1978). Efference copy and corollary discharge: implications for thinking and its disorders. *Schizophr. Bull.* 4, 636–640. doi: 10.1093/schbul/4.4.636

Feldman, H., and Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Front. Hum. Neurosci*. 4:215. doi: 10.3389/fnhum.2010.00215

Franck, N., Farrer, C., Georgieff, N., Marie-cardine, M., Dalery, J., d'Amato, T., et al. (2001). Defective recognition of one's own actions in schizophrenic patients. *Am. J. Psychiatry* 158, 454–459. doi: 10.1176/appi.ajp.158.3.454

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. Royal Soc. B* 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Friston, K. J. (2005). Hallucinations and perceptual inference. *Behav. Brain Sci.* 28, 764–766. doi: 10.1017/S0140525X05290131

Friston, K. J., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* 159, 417–458. doi: 10.1007/s11229-007-9237-y

Frith, C. (2005). The neural basis of hallucinations and delusions. *C.R. Biol.* 328, 169–175. doi: 10.1016/j.crvi.2004.10.012

Frith, C. (2012). Explaining delusions of control: the comparator model 20years on. *Conscious. Cogn*. 21, 52–54. doi: 10.1016/j.concog.2011.06.010

Frith, C. D., and Done, D. J. (1989). Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action. *Psychol. Med*. 19, 359–363. doi: 10.1017/S003329170001240X

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/S1364-6613(99)01417-5

Garbarini, F., Mastropasqua, A., Sigaudo, M., Rabuffetti, M., Piedimonte, A., Pia, L., et al. (2016). Abnormal sense of agency in patients with schizophrenia: evidence from bimanual coupling paradigm. *Front. Behav. Neurosci*. 10:43. doi: 10.3389/fnbeh.2016.00043

Garber, M. (2014). *When PARRY met ELIZA: A Ridiculous Chatbot Conversation From 1972*. The Atlantic.

Garety, P. A., and Freeman, D. (1999). Cognitive approaches to delusions: a critical review of theories and evidence. *British J. Clin. Psychol.* 38, 113–154. doi: 10.1348/014466599162700

Gould, L. N. (1950). Verbal hallucinations as automatic speech. The reactivation of dormant speech habit. *Am. J. Psychiatry* 107, 110–119. doi: 10.1176/ajp.107.2.110

Graham-Schmidt, K. T., Martin-Iverson, M. T., and Waters, F. A. V. (2018). Self- and other-agency in people with passivity (first rank) symptoms in schizophrenia. *Schizophr. Res*. 192, 75–81. doi: 10.1016/j.schres.2017.04.024

Haggard, P. (2017). Sense of agency in the human brain. *Nat. Rev. Neurosci.* 18, 197–208. doi: 10.1038/nrn.2017.14

Haggard, P., and Clark, S. (2003). Intentional action: conscious experience and neural prediction. *Conscious. Cogn*. 12, 695–707. doi: 10.1016/S1053-8100(03)00052-7

Haggard, P., Clark, S., and Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nat. Neurosci*. 5, 382–385. doi: 10.1038/nn827

Haggard, P., Martin, F., Taylor-clarke, M., Jeannerod, M., and Franck, N. (2003). Awareness of action in schizophrenia. *Neuroreport* 14, 1081–1085. doi: 10.1097/00001756-200305230-00035

Haggard, P., and Tsakiris, M. (2009). The experience of agency: feelings, judgments, and responsibility. *Curr. Dir. Psychol. Sci*. 18, 242–246. doi: 10.1111/j.1467-8721.2009.01644.x

Hauser, M., Knoblich, G., Repp, B. H., Lautenschlager, M., Gallinat, J., Heinz, A., et al. (2011a). Altered sense of agency in schizophrenia and the putative psychotic prodrome. *Psychiatry Res*. 186, 170–176. doi: 10.1016/j.psychres.2010.08.003

Hauser, M., Moore, J. W., de Millas, W., Gallinat, J., Heinz, A., Haggard, P., et al. (2011b). Sense of agency is altered in patients with a putative psychotic prodrome. *Schizophr. Res*. 126, 20–27. doi: 10.1016/j.schres.2010.10.031

Hoffman, R. E., Grasemann, U., Gueorguieva, R., Quinlan, D., Lane, D., and Miikulainen, R. (2011). Using computational patients to evaluate illness mechanisms in schizophrenia. *Biol. Psychiatry* 69, 997–1005. doi: 10.1016/j.biopsych.2010.12.036

Hughes, G. (2018). The role of the temporoparietal junction in implicit and explicit sense of agency. *Neuropsychologia* 113, 1–5. doi: 10.1016/j.neuropsychologia.2018.03.020

Hur, J., Kwon, J. S., Lee, T. Y., and Park, S. (2014). The crisis of minimal self-awareness in schizophrenia: a meta-analytic review. *Schizophr. Res.* 152, 58–64. doi: 10.1016/j.schres.2013.08.042

Idei, H., Murata, S., Chen, Y., Yamashita, Y., Tani, J., and Ogata, T. (2018). A neurorobotics simulation of autistic behavior induced by unusual sensory precision. *Computational Psychiatry* 2, 164–182. doi: 10.1162/cpsy_a_00019

Jardri, R., Pouchet, A., Pins, D., and Thomas, P. (2011). Cortical activations during auditory verbal hallucinations in schizophrenia: a coordinate-based meta-analysis. *Am. J. Psychiatry*. 168, 73–81. doi: 10.1176/appi.ajp.2010.09101522

Jones, S. R., and Fernyhough, C. (2007). Thought as action: inner speech, self-monitoring, and auditory verbal hallucinations. *Conscious. Cogn.* 16, 391–399. doi: 10.1016/j.concog.2005.12.003

Kahl, S., and Kopp, S. (2018). A predictive processing model of perception and action for self-other distinction. *Front. Psychol.* 9:2421. doi: 10.3389/fpsyg.2018.02421

Knoblich, G., Stottmeister, F., and Kircher, T. (2004). Self-monitoring in patients with schizophrenia. *Psychol. Med.* 34, 1561–1569. doi: 10.1017/S0033291704002454

Körding, K. P., and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature* 427, 244–247. doi: 10.1038/nature02169

Koreki, A., Maeda, T., Fukushima, H., Umeda, S., Takahata, K., Okimura, T., et al. (2015). Behavioral evidence of delayed prediction signals during agency attribution in patients with schizophrenia. *Psychiatry Res.* 230, 78–83. doi: 10.1016/j.psychres.2015.08.023

Kot, T., and Serper, M. (2002). Increased susceptibility to auditory conditioning in hallucinating schizophrenic patients: a preliminary investigation. *J. Nerv. Ment. Dis.* 190, 282–288. doi: 10.1097/00005053-200205000-00002

Krekelberg, B. (2010). Saccadic suppression. *Current Biol.* 20, 228–229. doi: 10.1016/j.cub.2009.12.018

Laroi, F., Luhrmann, T. M., Bell, V., Christian, W. A., Deshpande, S., Fernyhough, C., et al. (2014). Culture and hallucinations: overview and future directions. *Schizophr. Bull.* 40, 213–220. doi: 10.1093/schbul/sbu012

Legaspi, R., and Toyoizumi, T. (2019). A Bayesian psychophysics model of sense of agency. *Nat. Commun.* 10, 1–11. doi: 10.1038/s41467-019-12170-0

Leggenhager, B., Pazzaglia, M., Scivoletto, G., Molinari, M., and Aglioti, S. M. (2012). The sense of the body in individuals with spinal cord injury. *PLoS ONE* 7:e50757. doi: 10.1371/journal.pone.0050757

Lemaitre, A. L., Luyat, M., and Lafargue, G. (2016). Individuals with pronounced schizotypal traits are particularly successful in tickling themselves. *Conscious. Cogn.* 41, 64–71. doi: 10.1016/j.concog.2016.02.005

Lindner, A., Thier, P., Kircher, T. T. J., Haarmeier, T., and Leube, D. T. (2005). Disorders of agency in schizophrenia correlate with an inability to compensate for the sensory consequences of actions. *Curr. Biol.* 15, 1119–1124. doi: 10.1016/j.cub.2005.05.049

Maeda, T., Kato, M., Muramatsu, T., Iwashita, S., Mimura, M., and Kashima, H. (2012). Aberrant sense of agency in patients with schizophrenia: forward and backward over-attribution of temporal causality during intentional action. *Psychiatry Res.* 198, 1–6. doi: 10.1016/j.psychres.2011.10.021

Malenka, R., Angel, R. W., Hampton, B., and Berger, P. A. (1982). Impaired central error-correcting behavior in schizophrenia. *Arch. Gen. Psychiatry* 39, 101–107. doi: 10.1001/archpsyc.1982.04290010073013

Mamassian, P., and Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Res.* 38, 2817–2832. doi: 10.1016/S0042-6989(97)00438-0

Melcher, D. (2011). Visual stability. *Philos. Trans. Royal Soci. B* 366, 468–475. doi: 10.1098/rstb.2010.0277

Moore, J., and Haggard, P. (2008). Awareness of action: inference and prediction. *Conscious. Cogn.* 17, 136–144. doi: 10.1016/j.concog.2006.12.004

Moore, J., and Pope, A. (2014). The intentionality bias and schizotypy. *Q. J. Exp. Psychol.* 67, 2218–2224. doi: 10.1080/17470218.2014.911332

Moore, J. W. (2016). What is the sense of agency and why does it matter? *Front. Psychol.* 7:1272. doi: 10.3389/fpsyg.2016.01272

Moore, J. W., Dickinson, A., and Fletcher, P. C. (2011a). Sense of agency, associative learning, and schizotypy. *Conscious. Cogn.* 20, 792–800. doi: 10.1016/j.concog.2011.01.002

Moore, J. W., and Fletcher, P. C. (2012). Sense of agency in health and disease: a review of cue integration approaches. *Conscious. Cogn.* 21, 59–68. doi: 10.1016/j.concog.2011.08.010

Moore, J. W., and Obhi, S. S. (2012). Intentional binding and the sense of agency: a review. *Conscious. Cogn.* 21, 546–561. doi: 10.1016/j.concog.2011.12.002

Moore, J. W., Turner, D. C., Corlett, P. R., Arana, F. S., Morgan, H. L., Absalom, A. R., et al. (2011b). Ketamine administration in healthy volunteers reproduces aberrant agency experiences associated with schizophrenia. *Cogn. Neuropsychiatry* 16, 364–381. doi: 10.1080/13546805.2010.546074

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., and Friston, K. J. (2014). Bayesian inferences about the self (and others): a review. *Conscious. Cogn.* 25, 67–76. doi: 10.1016/j.concog.2014.01.009

Ohata, W., and Tani, J. (2020). *Investigation of Mulimodal and Agential Interactions in Human-Robot Imitation, Based on Frameworks of Predictive Coding and Active Inference*. ARxiv.

Pavlov, I. P. (1928). *Lectures on Conditioned Reflexes: Twenty-Five Years of Objective Study of the Higher Nervous Activity (behaviour) of Animals*. New York, NY: Liverwright Publishing Corporation. doi: 10.1037/11081-000

Powers, A.lbert R., Kelley, M., and Corlett, P. R. (2016). Hallucinations as top-down effects on perception. *Biol. Psychiatry* 1, 393–400. doi: 10.1016/j.bpsc.2016.04.003

Powers, A. R., Mathys, C., and Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science* 357, 596–600. doi: 10.1126/science.aan3458

Renes, R. A., Vermeulen, L., Kahn, R. S., Aarts, H., and van Haren, N. E. M. (2013). Abnormalities in the establishment of feeling of self-agency in schizophrenia. *Schizophr. Res.* 143, 50–54. doi: 10.1016/j.schres.2012.10.024

Rosset, E. (2008). It's no accident: our bias for intentional explanations. *Cognition* 108, 771–780. doi: 10.1016/j.cognition.2008.07.001

Schmack, K., Gòmez-Carrillo de Castro, A., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.-D., et al. (2013). Delusions and the role of beliefs in perceptual inference. *J. Neurosci.* 33, 13701–13712. doi: 10.1523/JNEUROSCI.1778-13.2013

Schmack, K., Rothkirch, M., Priller, J., and Sterzer, P. (2017). Enhanced predictive signalling in schizophrenia. *Hum. Brain Mapp.* 38, 1767–1779. doi: 10.1002/hbm.23480

Schwarz, K. A., Pfister, R., and Büchel, C. (2016). Rethinking explicit expectations: connecting placebos, social cognition, and contextual perception. *Trends Cogn. Sci.* 20, 469–480. doi: 10.1016/j.tics.2016.04.001

Seal, M. L., Aleman, A., and McGuire, P. K. (2004). Compelling imagery, unanticipated speech and deceptive memory: neurocognitive models of auditory verbal hallucinations in schizophrenia. *Cogn. Neuropsychiatry* 9, 43–72. doi: 10.1080/13546800344000156

Seashore, C. E. (1895). *Measurements of Illusions and Hallucinations in Normal Life. Studies from the Yale Psychological Laboratory*, 3.

Shergill, S., Samson, G., Bays, P. M., Frith, C. D., and Wolpert, D. M. (2005). Evidence for sensory prediction deficits in schizophrenia. *Am. J. Psychiatry* 162, 2384–2386. doi: 10.1176/appi.ajp.162.12.2384

Shergill, S. S., White, T. P., Joyce, D. W., Bays, P. M., Wolpert, D. M., and Frith, C. D. (2014). Functional magnetic resonance imaging of impaired sensory prediction in schizophrenia. *JAMA Psychiatry* 71, 28–35. doi: 10.1001/jamapsychiatry.2013.2974

Sidarus, N., Chambon, V., and Haggard, P. (2013). Priming of actions increases sense of control over unexpected outcomes. *Conscious. Cogn.* 22, 1403–1411. doi: 10.1016/j.concog.2013.09.008

Škodlar, B., Dernovšek, M. Z., and Kocmur, M. (2008). Psychopathology of schizophrenia in Ljubljana (Slovenia) from 1881 to 2000: changes in the content of delusions in schizophrenia patients related to various sociopolitical, technical and scientific changes. *Int. J. Social Psychiatry* 54, 101–111. doi: 10.1177/0020764007083875

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., et al. (2018). The predictive coding account of psychosis. *Biol. Psychiatry* 84, 634–643. doi: 10.1016/j.biopsych.2018.05.015

Strother, L., House, K. A., and Obhi, S. S. (2010). Subjective agency and awareness of shared actions. *Conscious. Cogn.* 19, 12–20. doi: 10.1016/j.concog.2009.12.007

Synofzik, M., Lindner, A., and Thier, P. (2008c). The cerebellum updates predictions about the visual consequences of one ' s behavior. *Current Biol.* 18, 814–818. doi: 10.1016/j.cub.2008.04.071

Synofzik, M., Thier, P., Leube, D. T., Schlotterbeck, P., and Lindner, A. (2010). Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions. *Brain* 133, 262–271. doi: 10.1093/brain/awp291

Synofzik, M., Vosgerau, G., and Newen, A. (2008a). Beyond the comparator model: a multifactorial two-step account of agency. *Conscious. Cogn.* 17, 219–239. doi: 10.1016/j.concog.2007.03.010

Synofzik, M., Vosgerau, G., and Newen, A. (2008b). I move, therefore i am: a new theoretical framework to investigate agency and ownership. *Conscious. Cogn.* 17, 411–424. doi: 10.1016/j.concog.2008.03.008

Teufel, C., Kingdon, A., Ingram, J. N., Wolpert, D. M., and Fletcher, P. C. (2010). Deficits in sensory prediction are related to delusional ideation in healthy individuals. *Neuropsychologia* 48, 4169–4172. doi: 10.1016/j.neuropsychologia.2010.10.024

Teufel, C., Subramaniam, N., Dobler, V., Perez, J., Finnemann, J., Mehta, P. R., et al. (2015). Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1–6. doi: 10.1073/pnas.1503916112

Teufel, C., Subramaniam, N., and Fletcher, P. C. (2013). The role of priors in bayesian models of perception. *Front. Comput. Neurosci.* 7:25. doi: 10.3389/fncom.2013.00025

Thakkar, K. N., Diwadkar, V. A., and Rolfs, M. (2017). Oculomotor prediction: a window into the psychotic mind. *Trends Cogn. Sci.* 21, 344–356. doi: 10.1016/j.tics.2017.02.001

Thakkar, K. N., Nichols, H. S., Mcintosh, L. G., and Park, S. (2011). Disturbances in body ownership in schizophrenia: evidence from the rubber hand illusion and case study of a spontaneous out-of-body experience. *PLoS ONE* 6:e27089. doi: 10.1371/journal.pone.0027089

Thakkar, K. N., and Rolfs, M. (2019). Disrupted corollary discharge in schizophrenia: evidence from the oculomotor system. *Biol. Psychiatry* 4, 1–9. doi: 10.1016/j.bpsc.2019.03.009

Tsakiris, M. (2017). The multisensory basis of the self: from body to identity to others. *Q. J. Exp. Psychol.* 70, 597–609. doi: 10.1080/17470218.2016.1181768

Tsakiris, M., and Haggard, P. (2005). The rubber hand illusion revisited: visuotactile integration and self-attribution. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 80–91. doi: 10.1037/0096-1523.31.1.80

Tsakiris, M., Haggard, P., Franck, N., Mainy, N., and Sirigu, A. (2005). A specific role for efferent information in self-recognition. *Cognition* 96, 215–231. doi: 10.1016/j.cognition.2004.08.002

Von Helmholtz, H. (1878). "The facts of perception," in *Selected Writings of Herman von Helmholtz*, ed R. Kahl (Weslyan University Press).

Von Helmholtz, H. (1866). *Concerning the Perceptions in General. Treatise on Physiological Optics Iii.* Leipzig.

Vosgerau, G., and Newen, A. (2007). Thoughts, motor actions, and the self. *Mind Language* 22, 22–43. doi: 10.1111/j.1468-0017.2006.00298.x

Voss, M., Chambon, V., Wenke, D., Kuhn, S., and Haggard, P. (2017). In and out of control: brain mechanisms linking fluency of action selection to self-agency in patients with schizophrenia. *Brain* 140, 2226–2239. doi: 10.1093/brain/awx136

Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A., and Haggard, P. (2010). Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. *Brain* 133, 3104–3112. doi: 10.1093/brain/awq152

Waters, F. A. V., and Badcock, J. C. (2010). First-rank symptoms in schizophrenia: reexamining mechanisms of self-recognition. *Schizophr. Bull.* 36, 510–517. doi: 10.1093/schbul/sbn112

Wegner, D. M., and Wheatley, T. (1999). Apparent mental causation: sources of the experience of will. *Am. Psychol.* 54, 480–492. doi: 10.1037/0003-066X.54.7.480

Weizenbaum, J. (1976). *Computing Power and Human Reason*. New York, NY: W. H. Freeman and Company.

Whitford, T. J., Mitchell, A. M., and Mannion, D. J. (2017). The ability to tickle oneself is associated with level of psychometric schizotypy in non-clinical individuals. *Conscious. Cogn.* 52, 93–103. doi: 10.1016/j.concog.2017.04.017

Wolpe, N., Haggard, P., Siebner, H. R., and Rowe, J. B. (2013). Cue integration and the perception of action in intentional binding. *Exp. Brain Res.* 229, 467–474. doi: 10.1007/s00221-013-3419-2

Wolpert, D., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science* 269, 1880–1882. doi: 10.1126/science.7569931

Wolpert, D.aniel M., and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nat. Neurosci.* 3, 1212–1217. doi: 10.1038/81497

Wolpert, D. M., and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks.* 11, 1317–1329. doi: 10.1016/S0893-6080(98)00066-5

Yamashita, Y., and Tani, J. (2012). Spontaneous prediction error generation in schizophrenia. *PLoS ONE.* 7:e37843. doi: 10.1371/journal.pone.0037843

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 03 September 2020
doi: 10.3389/frai.2020.00069

A Bayesian Account of Generalist and Specialist Formation Under the Active Inference Framework

Anthony G. Chen[1], *David Benrimoh*[2,3]*, *Thomas Parr*[3] *and Karl J. Friston*[3]

[1] Department of Physiology, McGill University, Montreal, QC, Canada, [2] Department of Psychiatry, McGill University, Montreal, QC, Canada, [3] The Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, London, United Kingdom

This paper offers a formal account of policy learning, or habitual behavioral optimization, under the framework of Active Inference. In this setting, habit formation becomes an autodidactic, experience-dependent process, based upon what the agent sees itself doing. We focus on the effect of environmental volatility on habit formation by simulating artificial agents operating in a partially observable Markov decision process. Specifically, we used a "two-step" maze paradigm, in which the agent has to decide whether to go left or right to secure a reward. We observe that in volatile environments with numerous reward locations, the agents learn to adopt a generalist strategy, never forming a strong habitual behavior for any preferred maze direction. Conversely, in conservative or static environments, agents adopt a specialist strategy; forming strong preferences for policies that result in approach to a small number of previously-observed reward locations. The pros and cons of the two strategies are tested and discussed. In general, specialization offers greater benefits, but only when contingencies are conserved over time. We consider the implications of this formal (Active Inference) account of policy learning for understanding the relationship between specialization and habit formation.

Keywords: Bayesian, active inference, generative model, preferences, predictive processing, learning strategies

OPEN ACCESS

Edited by:
Sriraam Natarajan,
The University of Texas at Dallas,
United States

Reviewed by:
Christopher L. Buckley,
University of Sussex, United Kingdom
Fabio Aurelio D'Asaro,
University of Naples Federico II, Italy

***Correspondence:**
David Benrimoh
david.benrimoh@mail.mcgill.ca

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

Received: *14 December 2019*
Accepted: *28 July 2020*
Published: *03 September 2020*

Citation:
Chen AG, Benrimoh D, Parr T and Friston KJ (2020) A Bayesian Account of Generalist and Specialist Formation Under the Active Inference Framework. Front. Artif. Intell. 3:69. doi: 10.3389/frai.2020.00069

INTRODUCTION

Any self-organizing system must adapt to its surroundings if it is to continue existing. On a broad timescale, population characteristics change to better fit the ecological niche, resulting in evolution and speciation (Futuyma and Moreno, 1988). On a shorter timescale, organisms adapt to better exploit their environment through the process of learning. The degree or rate of adaptation is also important. Depending on the environment around the organism, specialization into a specific niche or favoring a more generalist approach can offer distinct advantages and pitfalls (Van Tienderen, 1991). While adopting a single, automatic, behavioral strategy might be optimal for static environments—in which contingencies are conserved—creatures that find themselves in more variable or volatile environments should entertain a broader repertoire of plausible behaviors.

We focus upon adaptation on the shorter timescale in this paper, addressing the issue of behavioral specialization formally within a Markov decision process formulation of Active Inference (Friston et al., 2017). Active inference represents a principled framework in which to describe Bayes optimal behavior. It depends upon the notion that creatures use an internal (generative) model to explain sensory data, and that this model incorporates beliefs about "how

I will behave." Under Active Inference, learning describes the optimization of model parameters—updating one's generative model of the world such that one acts in a more advantageous way in a given environment (Friston et al., 2016). Existing work has focused upon how agents learn the (probabilistic) causal relationships between hidden states of the world that cause sensations which are sampled (Friston et al., 2016, 2017b; Bruineberg et al., 2018; Kaplan and Friston, 2018; Parr and Friston, 2018). In this paper, we extend this formalism to consider learning of policies.

While it is clear that well-functioning agents can update their understanding of the meaning of cues around them—in order to adaptively modulate their behavior—it is also clear that agents can form habitual behaviors. For example, in goal-directed vs. habitual accounts of decision making (Gläscher et al., 2010), agents can either employ an automatic response (e.g., go left because the reward is always on the left) or plan ahead using a model of the world. Habitual responses are less computationally costly than goal-oriented responses; making it desirable to trust habits when they have been historically beneficial (Graybiel, 2008; Keramati et al., 2011). This would explain the effect of practice—as we gain expertise in a given task, the time it takes to complete that task and the subjective experience of planning during the task diminishes, likely because we have learned enough about the structure of the task to discern and learn appropriate habits (Klapp, 1995).

How may our Active Inference agent learn and select habitual behaviors? To answer this question, we introduce a novel feature to the Active Inference framework; namely, the ability to update one's policy space. Technically, a prior probability is specified over a set of plausible policies, each of which represents a sequence of actions through time. Policy learning is the optimization of this probability distribution, and optimizing the structure of this distribution (i.e., "structure learning") through Bayesian model comparison. Habitual behavior may emerge through pruning implausible policies, and reducing the number of behaviors that an agent may engage in. If an agent can account for its behavior without calling on a given policy, it can be pruned, resulting in a reduced policy space, allowing agents to infer which policy it is pursuing more efficiently. Note that in Active Inference, agents have to infer the policy they are pursuing, where this inference is heavily biased by prior beliefs and preferences about the ultimate outcomes. We argue that pruning of redundant behavioral options can account for the phenomenon of specialization (behavior highly adapted to specific environments), and the accompanying loss of flexibility. In addition to introducing Bayesian model reduction for prior beliefs about policies, we consider its biological plausibility, and its relationship with processes that have been associated with structure learning (i.e., the removal of redundant model parameters). Finally, through the use of illustrative simulations, we show how optimizing model structure leads to useful policies, the adaption of an agent to its environment, the effect of the environment on learning and the costs and benefits of specialization. In what follows, we will briefly review the tenets of Active Inference, describe our simulation set up and then review the behavioral phenomenology in light of the questions posed above.

MATERIALS AND METHODS

Active Inference

Under Active Inference, agents act to minimize their variational free energy (Friston, 2012) and select actions that minimizes variational free energy expected following the action. This imperative formalizes the notion that an adaptive agent should act to avoid being in surprising states, should they wish to continue their existence. In this setting, free energy acts as an upper bound on surprise and expected free energy stands in for expected surprise or uncertainty. As an intuitive example, a human sitting comfortably at home should not expect to see an intruder in her kitchen, as this represents a challenge to her continued existence; as such, she will act to ensure that outcomes (i.e., whether or not an intruder is present) match her prior preferences (not being in the presence of an intruder); for example, by locking the door.

More formally, surprise is defined as the negative log probability of observed outcomes under the agent's internal model of the world, where outcomes are generated by hidden states (which the agents have no direct access to, but which cause the outcomes) that depend on the policies which the agent pursues (Parr and Friston, 2017):

$$-\ln P(\tilde{o}) = -\ln\left[\sum_{\tilde{s},\pi} P(\tilde{o},\tilde{s},\pi)\right] \tag{1}$$

Here, $\tilde{o} = (o_1, \ldots, o_T)$ and $\tilde{s} = (s_1, \ldots, s_T)$ correspond to outcomes (observations) and states throughout time, respectively, and π represents the policies (sequence of actions through time). Since the summation above is typically intractable, we can instead use free energy as an upper bound on surprise (Friston et al., 2017):

$$F = E_Q[\ln Q(\tilde{s},\pi) - \ln P(\tilde{o},\tilde{s},\pi)] \tag{2}$$

As an agent acts to minimize their free energy, they must also look forward in time and pursue the policy which they expect would best minimize their free energy. The contribution to the expected free energy from a given time, $G(\pi, \tau)$, is the free energy associated with that time, conditioned on the policy, and averaged with respect to a posterior predictive distribution (Friston et al., 2015):

$$G(\pi,\tau) = E_{Q(s_\tau|\pi)P(o_\tau|s_\tau)}\left[\ln P(o_\tau, s_\tau \mid \pi) - \ln Q(s_\tau \mid \pi)\right] \tag{3}$$

We can then sum over all future time-points (i.e., taking the path integral from the current to the final time: $(\pi) = \sum_{t\geq\tau} G(\pi,t)$) to arrive at the total expected free energy expected under each policy.

Partially Observable Markov Decision Process and the Generative Model

A Partially Observable Markov Decision Process (POMDP or MDP for short) is a generative model for modeling discrete

hidden states with probabilistic transitions that depend upon a policy. This framework is useful for formalizing planning and decision making problems and has various applications in artificial intelligence and robotics (Kaelbling et al., 1998). An MDP comprises two types of *hidden* variables which the agent must infer: hidden *states* ($\tilde{s}$) and *policies* (π). An MDP agent must then navigate its environment, armed with a generative model that specifies the joint probability distribution of observed outcomes and their hidden causes, and the imperative of minimizing free energy. The states, outcomes and policies are defined more concretely in the following sections.

The MDP implementation consists of the following matrices specifying categorical distributions (Friston et al., 2017b):

$$\begin{aligned} \mathbf{A_{ij}} &= P\left(o_\tau = i \mid s_\tau = j\right) && \text{state} - \text{outcome mapping} \\ \mathbf{B}\left(\mathbf{u}\right)_{\mathbf{ij}} &= P\left(s_{\tau+1} = i \mid s_\tau = j,\right. \\ & \left. u = \pi\left(\tau\right)\right) && \text{state} - \text{state transition} \\ \mathbf{C}_{\tau,\mathbf{i}} &= P\left(o_\tau = i\right) && \text{outcome preference} \\ \mathbf{D_i} &= P\left(s_1 = i\right) && \text{belief about initial states} \\ \mathbf{E_i} &= P\left(\pi = i \mid E\right) && \text{independent policy prior} \end{aligned}$$

The generative model (**Figure 1**) assumes that outcomes depend upon states, and that current states depend upon states at the previous timepoint and the action taken (as a result of the policy pursued). Specifically, the state-outcome relationship is captured by an **A** (likelihood) matrix, which maps the conditional probability of any *i-th* outcome given a *j-th* state. A policy, $\pi_i = (u_1, \ldots, u_T)$, is a sequence of actions (u) through time, which the agent can pursue. Generally, an agent is equipped with multiple policies it can pursue. Conceptually, these may be thought of as hypotheses about how to act. As hidden states are inaccessible, the agent must infer its current state from the (inferred) state it was previously in, as well as the policy it is pursuing. State-to-state transitions are described by the **B** (transition) matrix. The **C** matrix encodes prior beliefs about (i.e., a probability distribution over) outcomes, which are synonymous with the agent's preferences. This is because the agent wishes to minimize surprise and therefore will endeavor to attain outcomes that match the distributions in the **C** matrix. The **D** matrix is the prior belief about the agent's initial states (the agent's beliefs about where it starts off). Finally, **E** is a vector of the belief-independent prior over policies (i.e., intrinsic probability of each policy, without considering expected free energy).

A concept that will become important below is *ambiguity*. Assuming an agent is in the *i-th* hidden states, s^i, the probable outcomes are described by a categorical distribution by the *i-th* column of the **A** matrix. We can therefore imagine a scenario where the distribution $P\left(o_\tau \mid s_\tau = i\right)$ has *high entropy* (e.g., uniformly distributed), and outcomes are approximately equally likely to be sampled. This is an *ambiguous* outcome. On the other hand, we can have the opposite situation with an *unambiguous* outcome, where the distribution of outcomes given states has *low entropy*. In other words, "if I am in this state, then I will see this and only this." This unambiguous, precise outcome allows the agent to infer the hidden state that they are in.

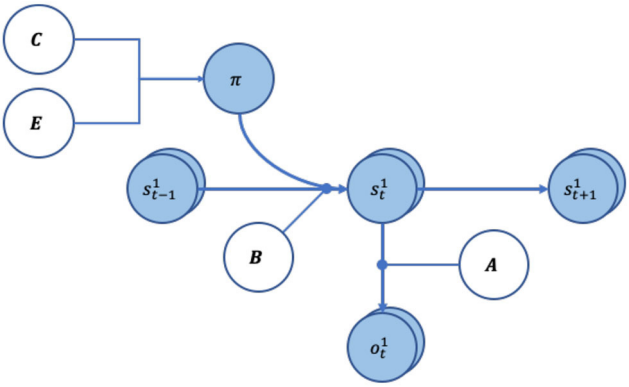


FIGURE 1 | Graphical representation of the generative model. The arrows indicate conditional dependencies, with the endpoint being dependant on where the arrow originated form. The variables in white circles show priors, whereas variables in light blue circles are random variables. The A and B matrices have round arrowhead to show they encode the transition probabilities between the variables.

Crucially, under Active Inference, an agent must also infer which policy it is pursuing at each time step. This is known as planning as inference (Botvinick and Toussaint, 2012). The requisite policy inference takes the form:

$$\pi = \sigma(\hat{\mathbf{E}} - \mathbf{F} - \gamma \cdot \mathbf{G}) \tag{4}$$

Here, π represents a vector of sufficient statistics of the posterior belief about policies: i.e., expectations that each allowable policy is currently in play. ***F*** is the free energy for each policy based on past time points and ***G*** is the expected free energy for future time points. The free energy scores the evidence that each policy is being pursued, while the expected free energy represents the prior belief that each policy will reduce expected surprise or uncertainty in the future. The expected free energy comprises two parts—*risk* and *ambiguity*. Risk is the difference between predicted and preferred outcomes, while ambiguity ensures that policies are chosen to disclose salient information. These two terms can be rearranged into *epistemic* and *pragmatic* components which, as one might guess, reduce uncertainty about hidden states of the world and maximize the probability of preferred outcomes.

The two quantities required to form posterior beliefs about the best policy (i.e., the free energy and expected free energy of each policy) can be computed using the **A**, **B**, and **C** matrices (Friston et al., 2016; Mirza et al., 2016). The variable γ is an inverse temperature (precision) term capturing confidence in policy selection, and $\hat{\mathbf{E}}$ is the (expected log of the) intrinsic prior probabilities in the absence of any inference (this is covered more in-depth in the "*Policy Learning and Dirichlet Parameters*" section below). The three quantities are passed through a softmax function (which normalizes the exponential of the values to sum to one). The result is the posterior expectation; namely, the most likely policy that the agent believes it is in. This expectation enables the agent to select the action that it thinks is most likely.

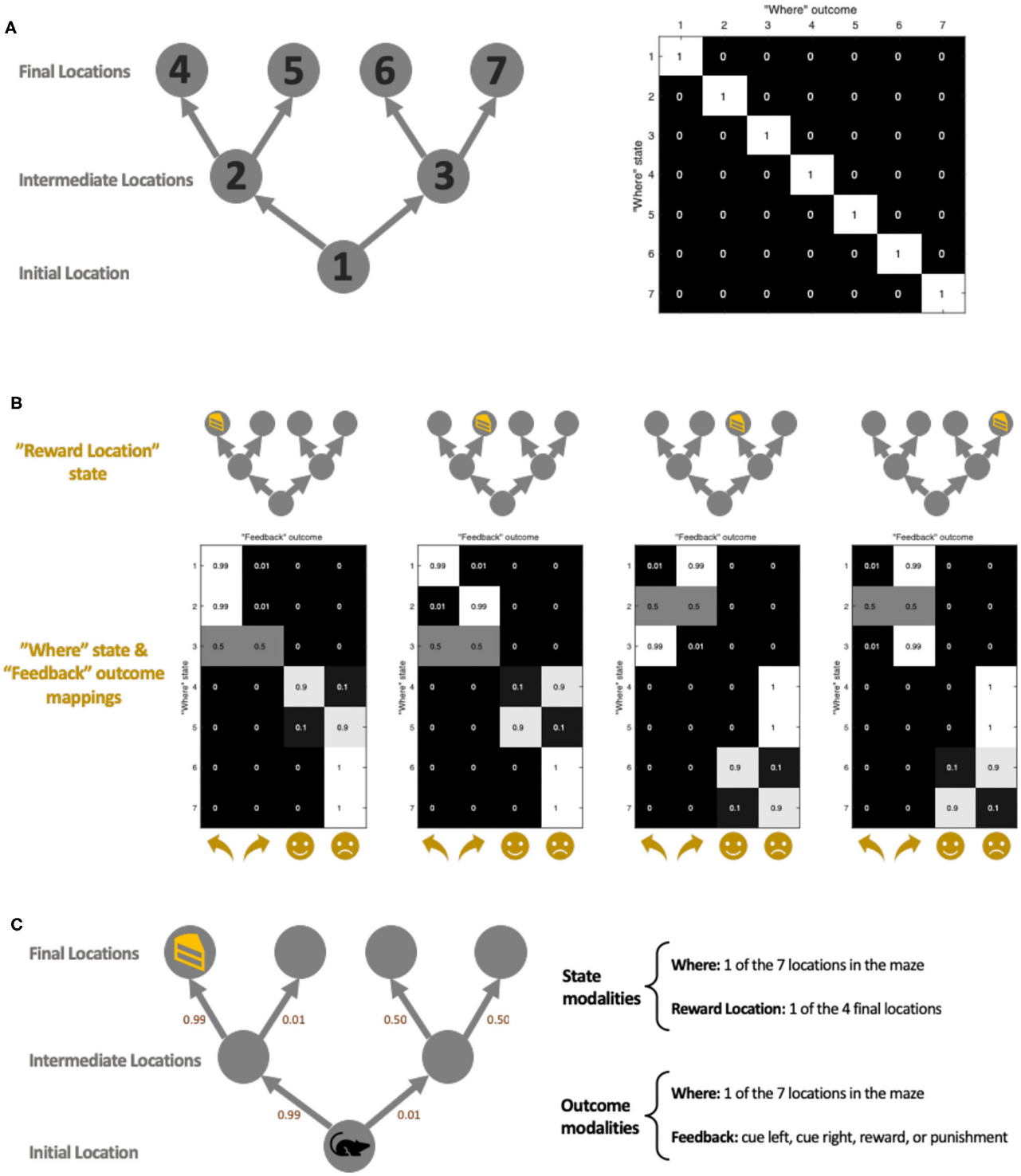


FIGURE 2 | Simulation maze set-up. **(A)** The maze location set-up. There are a total of 7 locations in the maze, each with their corresponding indexes (left diagram). The state-outcome mapping (A matrix) between "Where" (i.e., agent's current location) state and outcome is an identity matrix (right figure), meaning they always

(Continued)

FIGURE 2 | correspond exactly. The maze consists of three stages: initial, intermediate, and final. The state-state transition matrix (B matrix) ensures that an agent can only move forward in the maze, following the direction of the arrow. **(B)** The state-outcome transition probability between the "Where" state and "Feedback" outcome (as encoded by the A matrix). Depending on the location of the reward, the agent receives different feedbacks which include a directional cue (cue left or cue right) in the *initial* and *intermediate locations*, and a reward or punishment at the *final locations*. The index of the y-axis corresponds with the location index in **(A)**. Here we have depicted *unambiguous cues*, where the agent is 99% sure it sees the cue pointed in the correct (i.e., toward the reward location) cue. **(C)** An example maze set-up with a reward at the left-most *final location*. The agent starts in the *initial location*, and the agent's model-based brain contains representations of where it is in the maze, as well as where it thinks the reward is. The agent is able to make geographical observations to see where it is in the maze **(A)**, as well as receive a "feedback" outcome which gives it a cue to go a certain location, or to give it reward/punishment **(B)**. The small numbers beside each arrow illustrate the ambiguity of the cues. As an example, we have illustrated the left-most scenario of **(B)**.

Simulations and Task Set-Up

We return to our question of the effect of the environment on policy learning via setting up a simulated environment in which our synthetic agent (visualized as a mouse) forages (**Figures 2A,C**). Our environment takes the form of a two-step maze inspired by Daw et al. (2011), which is similar to that used in previous work on Active Inference (Friston et al., 2015, 2017). The maze allows for an array of possible policies, and the challenge for our agent is to learn to prioritize these appropriately. The agent has two sets of beliefs about the hidden states of the world: where it is in the maze, and where the reward is. The agent also receives two outcomes modalities: *where* it is in the maze and *feedback* received at each location in the maze (**Figure 2C**, right). The agent always knows exactly where it is in the maze (**Figure 2A**), and receives different "Feedback" outcomes, depending on where it is in the maze and the location of the reward (**Figure 2B**).

The mouse always starts in the same initial location (**Figure 2A**, position 1) and is given no prior information about the location of the reward. This is simulated by setting matrix **D** such that the mouse strongly believes that it is in the "initial location" at $\tau = 1$ but with a uniform distribution over the "reward location." The agent is endowed with a preference for rewarding outcomes and wishes to avoid punishing outcomes (encoded via the **C** matrix). Cues are placed in the initial and intermediate locations (cue left and cue right). While the agent has no preference for the cues *per se*, it can leverage the cue information to make informed decisions about which way to go to receive the reward. In other words, cues offer the opportunity to resolve uncertainty and therefore have salient or epistemic value. **Figure 2C** shows the reward in the left-most final location, accompanied by an *unambiguous cue*—the agent is 99% sure that "cue left" means that the reward is actually on the left. This leads it to the correct reward location. The nature of the maze is such that the agent cannot move backward; i.e., once it reaches the intermediate location it can no longer return to the initial location. Once the agent gets to the final location, it will receive either a reward (if it is at the reward location) or be punished.

To see the effect of training under different environments, we set up two different maze conditions: a *volatile environment*, in which the reward can appear in any one of the 4 final locations with equal frequencies, and a *non-volatile environment*, where the reward only appears on the two left final locations (**Figure 3A**). Crucially, this volatility is between-trial, because these contingencies do not change during the course of a trial. The mouse has no explicit beliefs about changes over multiple trials. Two mice with identical initial parameters are trained in these two distinct environments. With our set-up, each mouse can entertain 7 possible policies (**Figure 3B**). Four of the policies allow the mouse to get to one of the final four locations, whereas three additional policies result in the mouse staying in either the intermediate or initial locations. Finally, both mice are trained for 8 trials per day for 32 days with *unambiguous* cues in the two environments (**Figure 3C**). Bayesian model reduction (further discussed below) is performed in-between training to boost learning. Note that we set-up the training environment with *unambiguous cues* to allow for efficient learning, while the testing environment always has *ambiguous cues*—akin to explicit curriculums of school education vs. the uncertainty of real-life situations.

Policy Learning and Dirichlet Parameters

Whereas inference means optimizing expectations about hidden states given the current model parameters, learning is the optimization of the model parameters themselves (Friston et al., 2016). Within the MDP implementation of Active Inference, the parameters encode sets of categorical distributions that constitute the probabilistic mappings and prior beliefs denoted by **A**, **B**, **C, D,** and **E** above. A Dirichlet prior is placed over these distributions. Since the Dirichlet distribution is the conjugate prior for categorical distributions, we can update our Dirichlet prior with categorical data and arrive at a posterior that is still Dirichlet (FitzGerald et al., 2015).

While all model parameters can be learned (FitzGerald et al., 2015; Friston et al., 2016, 2017b), we focus upon policy learning. The priors are defined as follows:

$$E \sim Dir(e) \tag{5}$$

Here E is the Dirichlet distributed random variable (or parameter) that determines prior beliefs about policies. The variables $e = (e_1, \ldots, e_k)$ are the concentration parameters that parameterize the Dirichlet distribution itself. In the following, k is the number of policies. Policy learning occurs via the accumulation of e concentration parameters—the agent simply counts and aggregates the number of times it performs each policy and this count makes up the e parameters. Concretely, if we define $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ to be the probability the agent observes itself pursuing policies $\pi = 1, \ldots, k$, the posterior distribution over the policy space is:

$$Q(\mathbf{E}) = Dir(\mathbf{e}) = Dir(e + \pi) \tag{6}$$

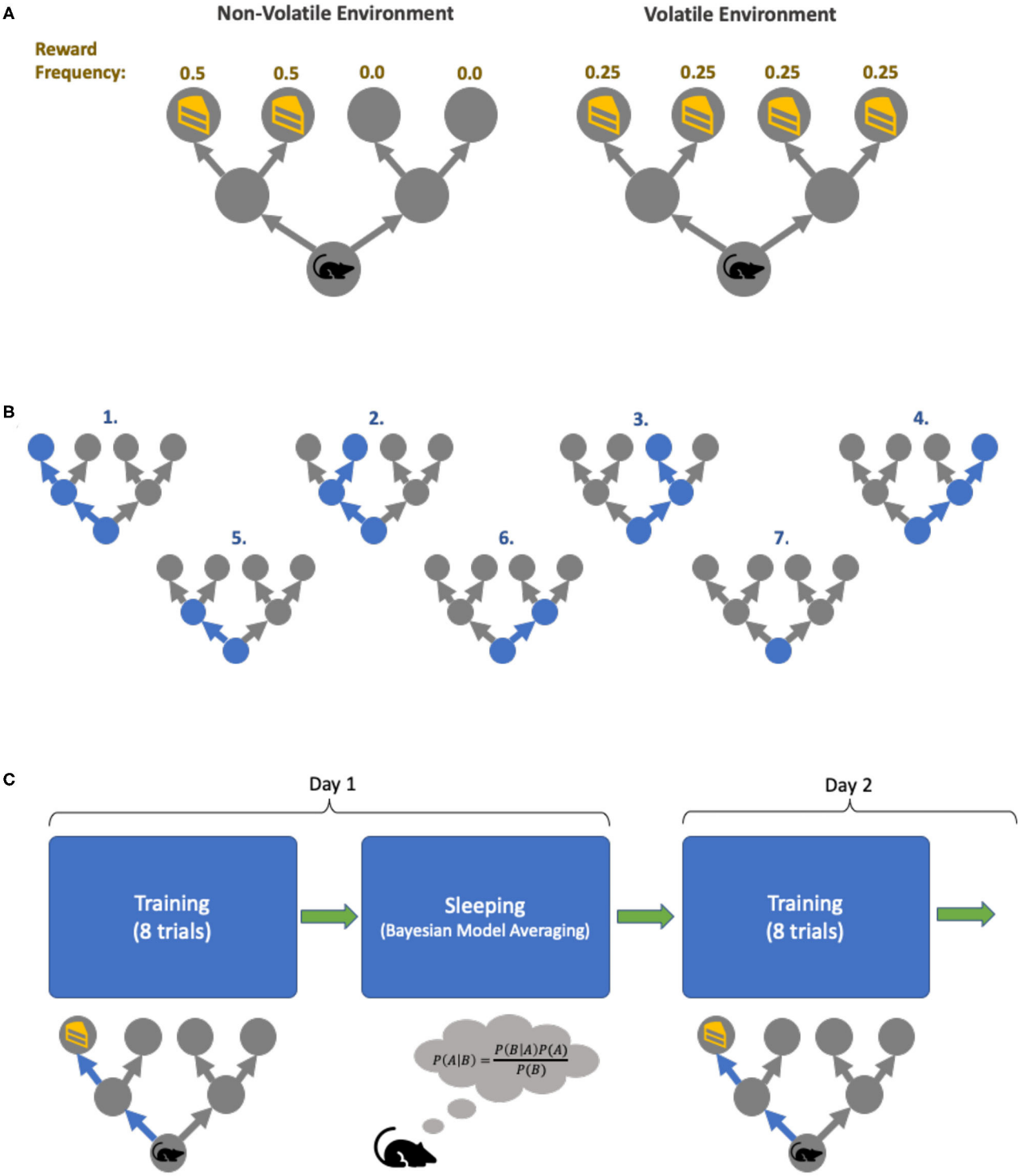


FIGURE 3 | Simulation task set-up. **(A)** The two environments in which the agents are trained. The environment can be *non-volatile* (left), in which the reward always appears on the *left* of the initial location, with equal frequency. The *volatile* environment (right) has reward appearing in all four final locations with equal frequencies. **(B)** The agent's policies. In our simulation, our agents each have 7 policies it can pursue: the first four policies correspond to the agent going to one of the final locations, policies 5–6 has the agent going to one of the intermediate locations and staying there, and policy 7 has the agent not moving from its initial location for the entire duration of a trial. **(C)** The training cycles. Each day, each agent is trained for 8 trials in their respective environment, and in between days the agent goes to "sleep" (and perform Bayesian model averaging to find more optimal policy concentrations). This process is repeated for 32 days.

where $\mathbf{e} = (e_1 + \pi_1, \ldots, e_k + \pi_k)$ is the posterior concentration parameter. In this way the Dirichlet concentration parameter is often referred to as a "pseudo-count." Intuitively, the higher the e parameter for a given policy, the more likely that policy becomes because more of $Q(\boldsymbol{E})$'s mass becomes concentrated around this policy. Finally, we take the expected logarithm to compute the posterior beliefs about policies in Equation (4):

$$\hat{\mathbf{E}} = \mathbb{E}_{Q(E)} \left[\ln P(\pi | E) \right] \tag{7}$$

The E vector can now be thought of as an empirical prior that accumulates the experience of policies that are carried over from previous trials. In short, it enables the agent to learn about the sorts of things that it does. This experience dependent prior policy enters inference via Equation (4). Before demonstrating this experience dependent learning, we look at another form of learning known variously as Bayesian model selection or structure learning.

Bayesian Model Comparison

In Bayesian model comparison, multiple competing hypotheses (i.e., models or the priors that defines models) are evaluated in relation to existing data and the model evidence for each is compared (Hoeting et al., 1999). Bayesian model averaging (BMA) enables one to use the results of Bayesian model comparison, by taking into account uncertainty about which is the best model. Instead of selecting just the most probable model, BMA allows us to weight models by their relative evidence—to evaluate model parameters that are a weighted average under each model considered. This is especially important in situations where there is no clear winning model (Hoeting et al., 1999).

An organism which harbors alternative models of the world needs to consider its own uncertainty about each model. The most obvious example of this is in the evaluation of different plausible courses of action (policies), each entailing a different sequence of transitions. Such models need to be learnt and optimized (Acuña and Schrater, 2010; FitzGerald et al., 2014) and, rejected, should they fall short. Bayesian model averaging is used implicitly in Active Inference when forming beliefs about hidden states of the world, where each policy is regarded as a model and different posterior beliefs about the trajectory of hidden states under each policy are combined using Bayesian model averaging. However, here, we will be concerned with the Bayesian model averaging over the policies themselves. In other words, the model in this instance becomes the repertoire of policies entertained by an agent.

Returning to our maze task, our artificial agents traverse through the maze each day and aggregate e parameters (Equation 6) to form its daily posterior—that will serve as tomorrow's empirical prior. During Bayesian model reduction, various reduced models are constructed, via strengthening and weakening amalgamations of e parameters. For each configuration of these policy parameters, model evidence is computed, and BMA performed to acquire the optimal posterior, which becomes the prior for the subsequent day. In brief, we evaluated the evidence of models in which each policy's prior concentration parameter was increased by eight, while the remainder were suppressed (by factor of two and four). This creates a model space—over which we can average to obtain the Bayesian model average of concentration parameters in a fast and biologically plausible fashion. Please see Appendix A, section A.1 for a general introduction to Bayesian model reduction and averaging. Appendix A, section A.2 provides an account of the procedures for an example "day." In what follows, we now look at the kinds of behaviors that emerge from day-to-day using this form of autodidactic policy learning—and its augmentation with Bayesian model averaging. We will focus on the behaviors that are elicited in the simulations, while the simulation details are provided in the appropriate figure legends (and open access software—see software note).

RESULTS

Learning

We now turn to our question about the effect of the environment on policy learning. Intuitively, useful policies should acquire a higher e concentration, becoming more likely to be pursued in the future. In simulations, one readily observes that policy learning occurs and is progressive, evident by the increase in e concentration for frequently pursued policies (**Figure 4**), which rapidly reach stable points within 10 days (**Figure 4B**, see **Figure 3C** for the concept of "training days"). Interestingly, the relative policy strengths attain stable points at different levels, depending on the environment in which the agent is trained. In a conservative environment, the two useful policies stabilize at high levels ($e \approx 32$), whereas in a volatile environment, these four useful policies do not reach the same accumulated strengths ($e \approx 25$). Furthermore, the policies that were infrequently used are maintained at lower levels when trained in a non-volatile environment ($e \approx 7$), while they are more likely to be considered for the agent trained in the volatile environment ($e \approx 11$).

We will henceforth refer to the agent trained in the non-volatile environment as the *specialist agent*, and the agent trained in the volatile environment as the *generalist agent.* Anthropomorphically, the specialist agent is, *a priori*, more confident about what to do: since the reward has appeared in the leftward location its entire life, it is confident that it will continue to appear in the left, thus it has predilections for left-going policies (policies 1 and 2 of **Figure 3B**). Conversely, the generalist agent has seen reward appear in multiple locations, thus it experiences a greater level of uncertainty and considers more policies as being useful, even the ones it never uses. We can think of these as being analogous to a general practitioner, who must entertain many possible treatment plans for each patient, compared to a surgeon who is highly skilled at a specific operation.

We can also illustrate the effect of training on the agents' *reward-acquisition rate:* the rate at which the agents successfully arrive at the reward location (**Figure 5**). Here, we tested the agents after each day's training. We see that (**Figure 5B**, left) with just a few days of training, the specialist agent learns the optimal policies and its *reward-acquisition rate* becomes consistently higher than a *naïve agent* with no preference over any of its policies ($e_{naive} = (e_1, \ldots, e_7) = (1, \ldots, 1)$). Conversely,

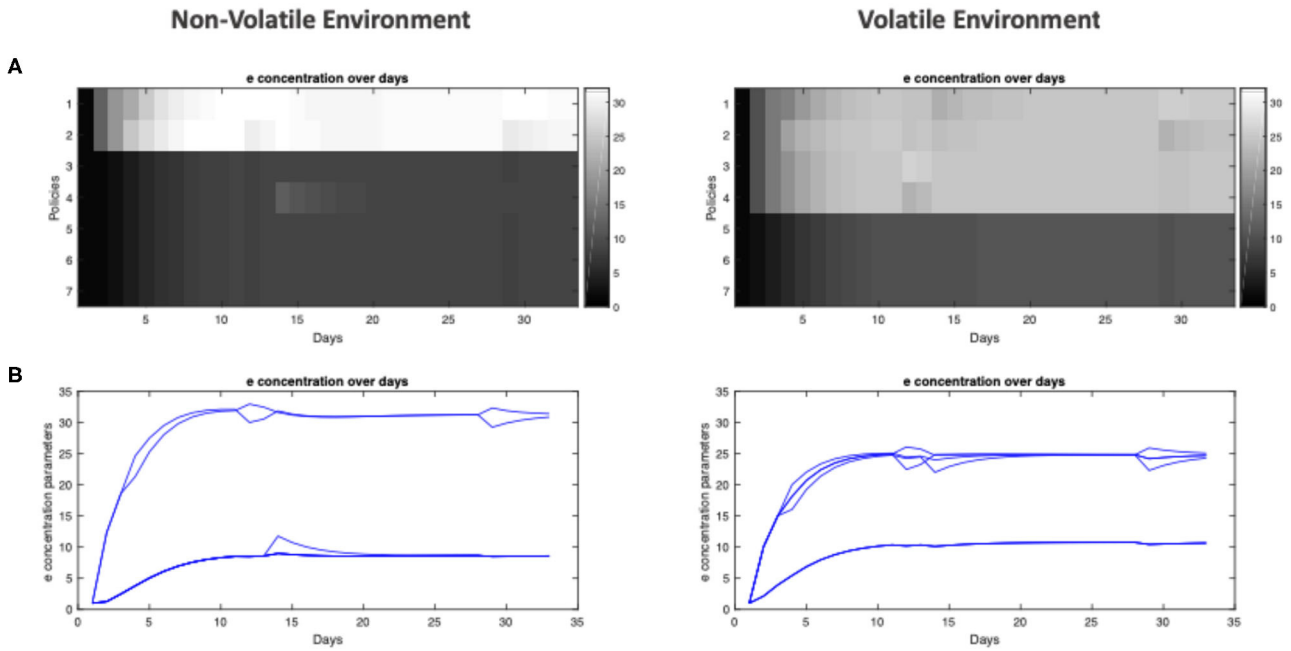


FIGURE 4 | Policy learning over days for agent training in non-volatile and volatile environments. **(A)** Heat-map of *e* concentration parameters for each policy (separated by rows) over all 32 days of training (separated by columns). **(B)** Plot of *e* concentration parameter for policies over 32 days of training.

the generalist agent never becomes an expert in traversing its environment. While it learns to identify the useful policies (**Figure 5A**, right), its performance is never significantly better than the naïve agent (**Figure 5B**, right). We emphasize that the "naive" agent does not simply select policies at random. Rather, it has uninformative policy priors and therefore relies upon its model-based component for policy inference (Equation 4). The similarity in performance between the generalist and "naive" agent is further discussed in the limitations section. Overall, we see that a *non-volatile* environment leads to specialization, whereas a *volatile* environment leads to the agent becoming a generalist.

Testing

We then asked how the specialist and generalist mice perform when transported to different environments. We constructed three testing environments (**Figure 6A**): the *specialized environment,* similar to the environment the specialized agent is trained in; namely, with rewards that only appear on the left side of the starting location (low volatility); the *general environment* containing rewards that may appear in any of the four final locations (high volatility); additionally, the *novel environment* has reward *only* on the right side of the starting location (low volatility).

Each agent was tested for 512 trials in each test environment. Note that the agents do not learn during the testing phase—we simply reset the parameters in our synthetic agents after each testing trial to generate perfect replications of our test settings. We observe that an untrained (naïve) agent has a baseline reward-acquisition rate of ∼60%. On the contrary, the specialist agent excels when the environment is similar to that it trained in, performing at the highest level (89%) out all the agents. In contrast, the specialist agent performs poorly in a general environment (46% reward-acquisition), and fails all but one out of its 512 attempts in a novel environment where it needs to go in the opposite direction to that of its training (**Figures 6B,C**). The generalist agent, being equally trained in all four policies—that take it to one of the end locations—does not suffer from reduced reward-acquisition when exposed to a new environment (the specialized environment or novel environment). However, it does not perform better in a familiar, general environment either. The agent's reward-acquisition remains around 60% across all testing environments, similar to that of a naïve agent (**Figures 6B,C**).

Overall, we find that becoming a specialist vs. a generalist has sensible trade-offs. The benefit of specialization is substantial when operating within the same environment, consistent with data on this topic in a healthcare setting (Harrold et al., 1999; Wu et al., 2001). However, if the underlying environment is different, then performances can decrease to one which is poorer than the performance without specialization.

DISCUSSION

Specialists and Generalists

Our focus in this paper has been on policy optimization, where discrete policies are optimized through learning and Bayesian model reduction. By simulating the development of specialism and generalism, we illustrated the capacity of a generalist to perform in a novel environment, but its failure to reach the level of performance of a specialist in a specific environment. We

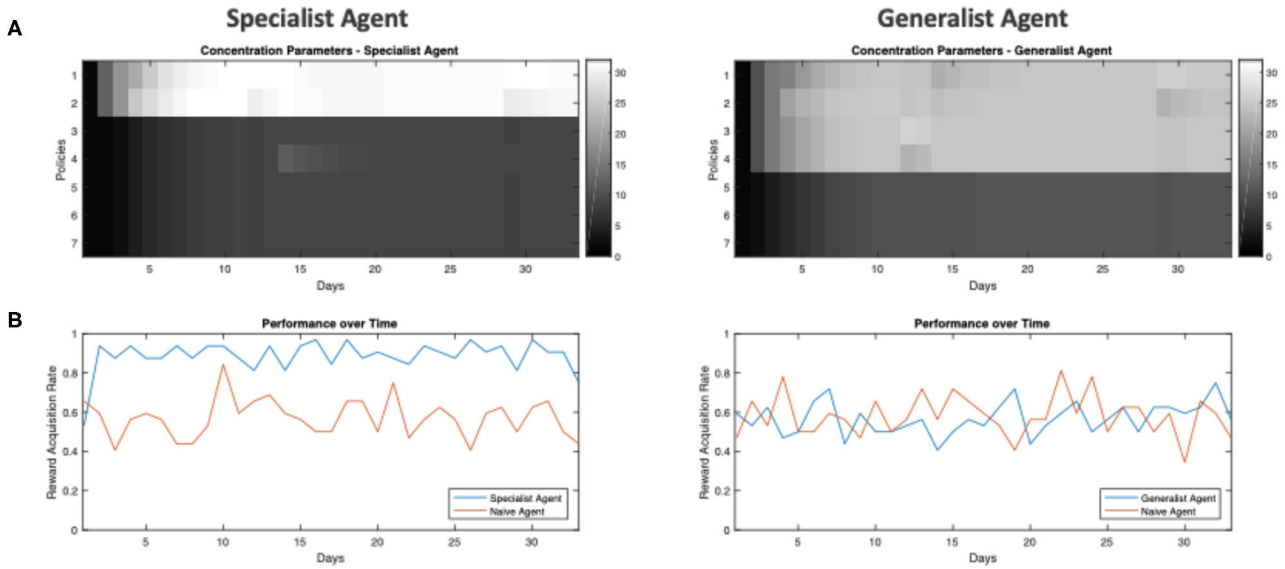


FIGURE 5 | Example performance of in-training agents over days. **(A)** Heat-map of *e* concentration parameters for each policy (separated by rows) over all 32 days of training (separated by columns). **(B)** The frequency at which the agent is able to get to the reward location when tested under ambiguity. This simulated testing is done after each day of training, where each agent is tested under ambiguity (the agent is 65% sure it sees the correct cue) for 32 trials, where the reward location / frequency in the testing environment is identical to the environment in which the agent is trained (i.e., a specialist agent is tested in an environment with low volatility and the reward always being on the left of the initial location). The frequency is computed from how many out of the 32 trials the agent is able to get to the true reward location.

now turn to a discussion of the benefits and costs of expertise. Principally, the drive toward specialization (or expertise) is the result of the organism's imperative to minimize free energy. As free energy is an upper bound on surprise (negative Bayesian model evidence), minimizing free energy maximizes model evidence (Friston et al., 2013). As model evidence takes into account both the accuracy and complexity of an explanation (FitzGerald et al., 2014), it is clear that having a parsimonious model that is well-suited to the environment—a specialist model—will tend to minimize free energy over time, provided the environment does not change.

In a stable (conservative, non-volatile) setting, a complex environment can be distilled down into a simple model without sacrificing accuracy. This results in efficient policy selection and provides a theoretical framework for understanding the formation of expertise. In our simulations, the agent trained in the unchanging environment learns to favor the two policies that go left, as the reward is always on the left of the starting location. It thus becomes more efficient and acts optimally in the face of uncertainty. This is evident by its excellent performance in finding left-situated rewards (**Figure 6**). Indeed, previous theories of expertise differentiate experts from novices in their ability to efficiently generate complex responses to their domain-specific situations (Krampe, 2002; Ericsson, 2008; Furuya and Kinoshita, 2008). For example, in typists, expertise is most well-characterized by the ability to quickly type different letters in succession using different hands (Gentner, 1998; Krampe, 2002). In essence, the expert needs to quickly select from her repertoire of motor policies the most appropriate to type the desired word. This is a non-trivial problem: using just the English alphabet, there are a total of 26^m ways of typing an m-character-long word (e.g., a typist needs to select from $26^6 = 308915776$ policies to type the 6-letter word "EXPERT"). It is no wonder that a beginner typist struggles greatly and needs to forage for information by visually searching the keyboard for the next character after each keystroke. The expert, on the other hand, has an optimized prior over her policy space, and thus is able to efficiently select the correct policies to generate the correct character sequences.

However, specialization does not come without its costs. The price of expertise is reduced flexibility when adapting to new environments, especially when the new settings are contradictory to previous settings (Sternberg and Frensch, 1992; Graybiel, 2008). Theoretically, the expert has a simplified model of their domain, and, throughout their extensive training, has the minimum number of parameters necessary to maintain their model's high accuracy. Consequently, it becomes difficult to fit this model to data in a new, contradictory environment that deviates significantly from the expert's experience. For instance, we observe that people trained in a perceptual learning task perform well in the same task, but perform worse than naïve subjects when the distractor and target set are reversed—and take much longer to re-learn the optimal response than new subjects who were untrained (Shiffrin and Schneider, 1977).

Conversely, a volatile environment precludes specialization. The agent cannot single-mindedly pursue mastery in any particular subset of policies, as doing so would come at the cost of reduced accuracy (and an increase in free energy). The generalist agent therefore never reaches the level of performance that the specialist agent is capable of at its best. Instead, the generalist performs barely above the naïve average reward-acquisition

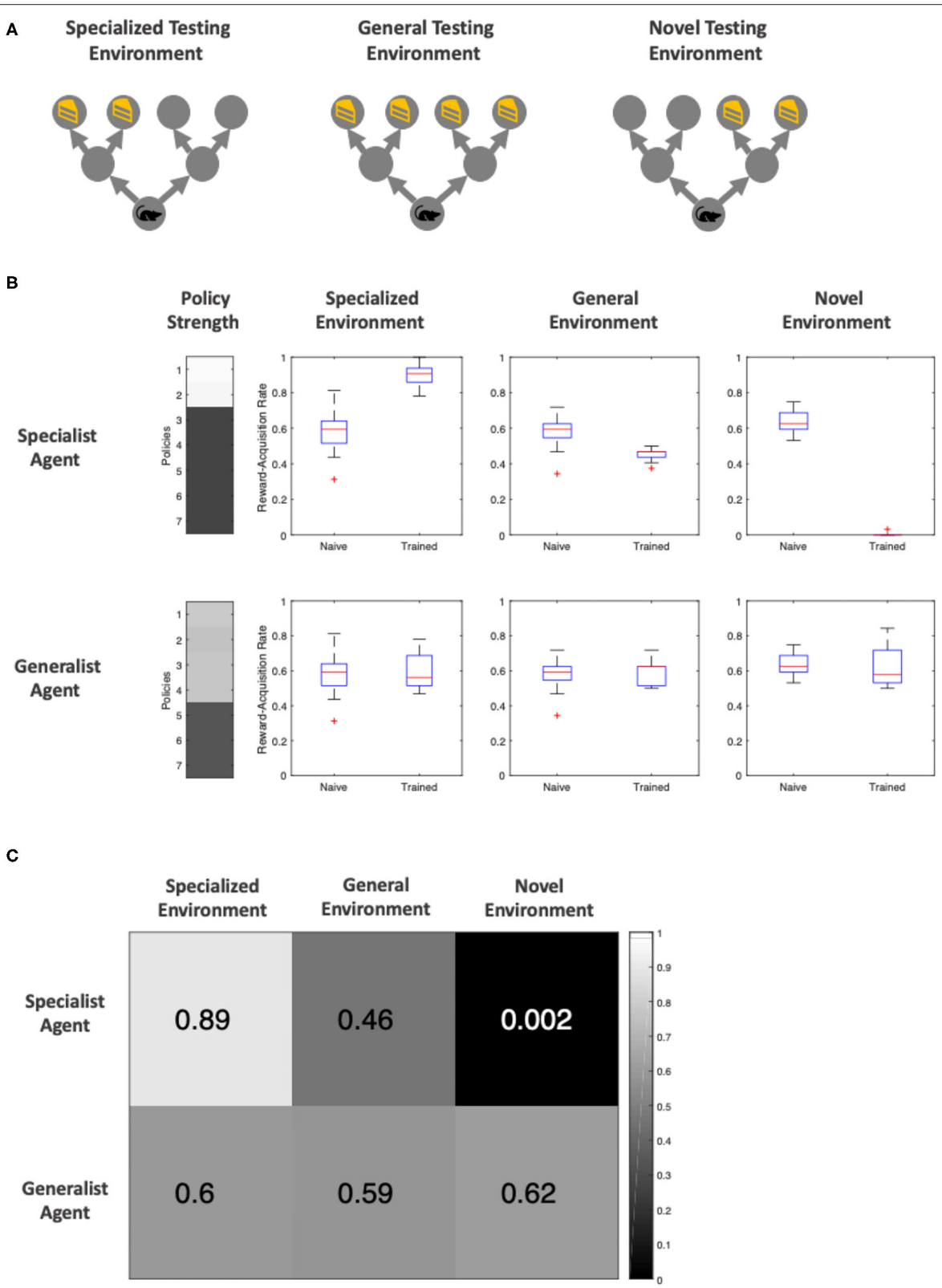


FIGURE 6 | Post-training performance of specialist and generalist agents in ambiguous environments (the agent is 65% sure it sees the cue telling it to go in the correct direction) **(A)** Visualization of the three testing environments. The *specialized* and *general* testing environment have identical reward location and frequencies

(Continued)

FIGURE 6 | top the environments in which the *specialist* and *generalist* agents were trained, respectively. The novel environment is a new, low volatility environment in which the reward only appears to the *right* of the initial location. **(B)** Distribution of reward-acquisition-rate of specialist and generalist agents compared against a naïve agent with no training. The "Policy Strength" column shows how much of each policy the agent has learned, and the three boxes of boxplots show the comparison in performance. The reward-acquisition rate distribution is generated via running each trial 32 times to generate a reward-acquisition rate (proportion of times the agent correctly navigates to the reward location), and repeating this process 16 times to generate a distribution of scores. **(C)** A confusion matrix of mean reward-rate of each agent within each testing environment. Both the heat map and the color over each element represents the reward-acquisition rate.

rate, even when tested under a general environment. However, the generalist is flexible. When placed in novel and changing environments, it performs much better than our specialist agent.

Interestingly, we note that specialist formation requiring a *conservative* training environment adheres to the requirements specified by K. Anders Ericsson in his theory of *deliberate practice*—a framework for any individual to continuously improve until achieving mastery in a particular field (Ericsson et al., 1993, 2009; Ericsson, 2008). Ericsson establishes that deliberate practice requires a well-defined goal with clear feedback (c.f., low volatility learning environment) and ample opportunity for repetition and refinement of one's performance (c.f., training, repetition and, potentially, Bayesian model reduction).

While outside of the current scope, future work could consider even more dynamic (and potentially more realistic) situations where the goal changes intermittently. We tentatively predict if the agent is given time in environments where state-outcome mappings can be inferred easily (unambiguous), it will perform well irrespective of goal location. However, if the environment is always ambiguous, it will be more difficult to learn good habits, and even harder so with an itinerant goal.

Ways of Learning

There are two principal modes of (policy) learning. The first is *learning via reduction,* which entails a naïve agent that starts with an over-complete repertoire of possible policies, who then learns to discard the policies that are not useful. This is how we have tackled policy learning here; specifically, via optimizing a Dirichlet distribution over policies, using Bayesian model reduction. By starting with an abundance of possible policies, we ensure that the best policy is likely to always be present. This also corresponds with the neurobiological findings of childhood peaks in gray matter volume and number of synapses, followed by adolescent decline (Huttenlocher et al., 1982; Huttenlocher and Dabholkar, 1997; Giedd, 2008). In this conceptualization, as children learn they prune away redundant connections, much as our agents triage away redundant policies. Likewise, as the policy spaces are reduced and made more efficient, we also observe a corresponding adolescent decline in brain glucose usage (Chugani et al., 1987). This is consistent with the idea that informational complexity is metabolically more expensive (Landauer, 1961).

The second method of learning is *learning via expansion*. Here, we start with a very simple model and increase its complexity until a more optimal model is reached. Concretely, this problem of increasing a parameter space is one addressed by Bayesian Non-parametric modeling (Ghahramani, 2013), and has been theorized to be utilized biologically for structure learning to infer hidden states and the underlying structures of particular situations (Gershman and Niv, 2010; Collins and Frank, 2013).

Bayesian Model Comparison

In our simulations, we optimized policy strengths through the process of Bayesian model reduction (to evaluate the free energy or model evidence of each reduced model), followed by model averaging—in which we take the weighted average over *all* reduced models. However, BMA is just one way of using model evidences to form a new model. Here, we discuss other approaches to model comparison, their pros and cons, and biological implications. The first is Bayesian model *selection*, in which only the reduced model with the greatest evidence is selected to be the prior for the future, without consideration of competing models. This offers the advantage of reduced computational cost (no need to take the weighted sum during the averaging process) at the cost of a myopic selection—the uncertainty over reduced models is not taken into account.

The second method, which strikes a balance between BMA and Bayesian model selection with respect to the consideration of uncertainty, is BMA with *Occam's Window* (Raftery, 1995). In short, a threshold is established, O_R, and if the log evidence of any reduced model is not within O_R, we simply do not consider that reduced model. Neurobiologically, this would correspond to the effective silencing of a synapse if it falls below a certain strength (Fernando et al., 2012). This way, multiple reduced models and relative uncertainties are still considered, but a great degree of computational cost is saved since less reduced models are considered overall.

We note that in Bayesian model comparison, the repertoire of reduced models to be considered, the width of the Occam's window, as well as the time spent in "wake" (experience-gathering) and "rest" (model comparison and reduction) phases are all hyperparameters. Similar to model parameters, we can expect there to be hyperpriors, which are priors over the hyperparameters. While outside of the scope of the current work, hyperpriors may be optimized via evolutionary processes which also reduce the (path integral of) free energy (Kirchhoff et al., 2018; Linson et al., 2018).

Furthermore, we theorize that there may be a connection between these model optimization processes, and those thought to occur during sleep, in line with previously theorized role of sleep in minimizing model complexity (Hobson and Friston, 2012), and related to the homeostasis hypothesis of sleep (Tononi and Cirelli, 2006). In this theory, a variational free energy minimizing creature tries to optimize a generative model that is both accurate and simple—i.e., that affords the least complicated explanation for the greatest number of

observations. Mathematically, this follows from the fact that surprise can be expressed as model evidence—and model evidence is the difference between accuracy and complexity. During wakefulness, an organism constantly receives sensory information, and forms accurate yet potentially complex models to fit these data (neurobiologically, via increases in the number and strength of synaptic connections through associative plasticity). During sleep, which lacks any precise sensory input, creatures can optimize their models *post-hoc* by reducing complexity (Friston and Penny, 2011). This can be achieved by considering reduced (simpler) models and seeing how well they explain the data collected during waking hours (FitzGerald et al., 2014). This is sometimes called Bayesian model reduction (Friston et al., 2018). While we refer to model reduction as "sleep" in this work, we acknowledge that no consensus has been reached on the role of sleep, and the function of sleep as Bayesian model reduction is just one theory.

Computational Psychiatry

Previously, Active Inference has been used as a tool for computational psychiatry, both for phenotyping (Schwartenbeck and Friston, 2016), and as a model of psychiatric symptoms such as illusions (Brown et al., 2013), visual hemineglect (Parr and Friston, 2018), and auditory hallucinations (Benrimoh et al., 2018), to name a few. For instance, low precision assigned to sensory attenuation can result in hallucination (Brown et al., 2013). Uniquely, Active Inference allows for the consideration of both perception and action. Specifically, some recent works have begun to show the potential for disruptions of the policy space to engender symptoms such as visual neglect (Parr and Friston, 2018) and auditory hallucination in schizophrenia (Benrimoh et al., 2018).

While the role of the policy space has been shown to be important, so far, there has been no formal account in Active Inference on how a policy space is learned—in the sense of structure learning—and altered. This is what the current work seeks to provide. Specifically, we formalize the policy to incorporate a policy prior. We then show how this prior is learned, as well as introducing the notion of Bayesian model reduction to change the structure of the policy space. Further, we showcase the interplay between the prior and the free energy in our "two-step" task, where we identified ambiguity—in the state-outcome mapping—as a crucial determinant of when policy priors (i.e., "habits") become important. Depending on the training environment, we demonstrate that different policy priors can underwrite sensible behavior.

Simply put, while we had known that disruption to the policy space plays a role in various psychiatric symptoms, we are now equipped with a formalism to tackle how the policy space can become maladapted to its environment. This can be an experience-dependent process, where rare policies with low priors are never considered. This may also be a result of model-comparison, where the models compared may not have full support over the policy space, or the model averaging process may not consider the full set of possible policies (e.g., due to computational constraints). These are tentative hypotheses, which future work can explore in greater depth.

Moreover, we have focused on ease-of-interpretability in this work and hope this paper can also act as a foundational "tutorial" for future work in Active Inference that seeks to investigate the interaction between the policy space and behavior. We have therefore refrained from making claims about specific brain areas. One can note that policies are usually associated with the striatum (Parr and Friston, 2018), while observation space is modality dependent, per the functional anatomy of primary and secondary sensory cortex (for instance, the state-outcome mapping in auditory tasks can be tentatively theorized to map to the Wernicke's—prefrontal connection). For more precise process theories on how the Active Inference machinery maps onto brain areas, we invite the readers to look at the discussion sections of Benrimoh et al. (2018) and Parr and Friston (2018).

Limitations

One limitation of our simulations was that our agents did not learn about cues at the same time they were learning about policies; in fact, the agents were constructed with priors on which actions were likely to lead to rewards, given specific cues (that is, a correctly perceived cue-left was believed by the agents to—and actually did—always lead to a reward on the left). As such, we did not model the learning of cue-outcome associations and how these may interact with habit formation. We argue this is a reasonable approximation to real behavior; where an animal or human first learns how cues are related to outcomes, and, once they have correctly derived a model of environmental contingencies, can then proceed to optimizing policy selection.

Additionally, while we were able to see a significant performance difference between specialist and generalist agents, there was little distinction between the performance of generalist and naïve agents. This likely resulted from the "two-step" maze being a relatively simple task. As agents are incentivized to go to the very end of the maze to receive a reward, the naïve agents and the generalist agents (as a result of the volatile training environment) have isomorphic prior beliefs about the final reward locations, and thus perform similarly. In this sense, becoming a generalist is the process of resisting specialization, and the preservation of naivety.

To address the above limitations, future work could involve more complex tasks to more clearly differentiate between specialist, generalist, and naïve agents. Additional types of learning should also be included, such as the learning of state-outcome mappings [optimizing the model parameters of the likelihood (**A**) matrix, as described in Friston et al. (2016, 2017b)], to understand how learning of different contingencies influence one another. In addition, more complex tasks may afford the opportunity to examine the generalization of specialist knowledge to new domains (Barnett and Ceci, 2002). This topic has recently attracted a great deal of attention from the artificial intelligence community (Pan and Yang, 2010; Hassabis et al., 2017).

Furthermore, it would be interesting to look at policy learning using a hierarchical generative model, as considered for deep temporal models (Friston et al., 2017a). This likely leads to a more accurate account of expertise-formation, as familiarity with a domain-specific task should occur at multiple-levels of the

neural-computation hierarchy (e.g., from lower level "muscle memory" to higher level planning). Likewise, more unique cases of learning can also be explored, such as the ability and flexibility to re-learn different tasks after specializing, and different ways of conducting model comparison (as discussed above).

CONCLUSION

In conclusion, we have presented a computational model under the theoretical framework of Active Inference that equips an agent with the machinery to learn habitual policies via a prior probability distribution over its policy space. In our simulations, we found that agents who specialize—employing a restricted set of policies because these were adaptive in their training environment—can perform well under ambiguity but only if the environment is similar to its training experiences. On the contrary, a generalist agent can more easily adapt to changing, ambiguous environments, but is never as successful as a specialist agent in a conservative environment. These findings cohere with the previous literature on expertise formation—as well as with common human experience. Finally, these findings may be important in understanding aberrant inference and learning in neuropsychiatric diseases.

DATA AVAILABILITY STATEMENT

All simulation scripts used for this article can be found on GitHub (https://github.com/im-ant/ActiveInference_PolicyLearning). Simulation is constructed using the MATLAB package SPM12 (https://www.fil.ion.ucl.ac.uk/spm/). Specifically, the DEM toolbox in SPM12 is used to run the Active Inference simulations.

AUTHOR CONTRIBUTIONS

DB, TP, and KF conceptualized the project and helped supervise it. AC was involved in the investigation along with DB. AC did the data curation, formal analysis, visualization, and writing of the first draft, while receiving methodology help from TP and KF. AC, DB, TP, and KF took part in the review and edits of the subsequent drafts. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Rosetrees Trust (Award Number 173346) to TP. KF was a Wellcome Principal Research Fellow (Ref: 088130/Z/09/Z).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at Chen et al. (2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00069/full#supplementary-material

REFERENCES

Acuña, D. E., and Schrater, P. (2010). Structure learning in human sequential decision-making. *PLoS Comput. Biol.* 6:e1001003. doi: 10.1371/journal.pcbi.1001003

Barnett, S. M., and Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychol. Bull.* 128, 612–637. doi: 10.1037/0033-2909.128.4.612

Benrimoh, D., Parr, T., Vincent, P., Adams, R. A., and Friston, K. (2018). Active inference and auditory hallucinations. *Comput. Psychiatry*. 2, 183–204. doi: 10.1162/cpsy_a_00022

Botvinick, M., and Toussaint, M. (2012). Planning as inference. *Trends Cogn. Sci.* 16, 485–488. doi: 10.1016/j.tics.2012.08.006

Brown, H., Adams, R. A, and Parees, I., Edwards, M., and Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognit. Process.* 14, 411–427.

Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., and Friston, K. J. (2018). Free-energy minimization in joint agent-environment systems: a niche construction perspective. *J. Theor. Biol.* 455, 161–178. doi: 10.1016/j.jtbi.2018.07.002

Chen, A. G., Benrimoh, D., Parr, T., and Friston, K. J. (2019). A Bayesian account of generalist and specialist formation under the active inference framework. *bioRxiv [Preprint]*. doi: 10.1101/644807

Chugani, H. T., Phelps, M. E., and Mazziotta, J. C. (1987). Positron emission tomography study of human brain functional development. *Ann. Neurol.* 22, 487–497. doi: 10.1002/ana.410220408

Collins, A. G. E., and Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychol. Rev.* 120, 190–229. doi: 10.1037/a0030852

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69, 1204–1215. doi: 10.1016/j.neuron.2011.02.027

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: a general overview. *Acad. Emerg. Med.* 15, 988–994. doi: 10.1111/j.1553-2712.2008.00227.x

Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev.* 100, 363–406. doi: 10.1037/0033-295X.100.3.363

Ericsson, K. A., Nandagopal, K., and Roring, R. W. (2009). Toward a science of exceptional achievement: attaining superior performance through deliberate practice. *Ann. N. Y. Acad. Sci.* 1172, 199–217. doi: 10.1196/annals.1393.001

Fernando, C., Szathmáry, E., and Husbands, P. (2012). Selectionist and evolutionary approaches to brain function: a critical appraisal. *Front. Comput. Neurosci.* 6:24. doi: 10.3389/fncom.2012.00024

FitzGerald, T. H. B., Dolan, R. J., and Friston, K. (2015). Dopamine, reward learning, and active inference. *Front. Comput. Neurosci.* 9:136. doi: 10.3389/fncom.2015.00136

FitzGerald, T. H. B., Dolan, R. J., and Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Front. Hum. Neurosci.* 8, 1–11. doi: 10.3389/fnhum.2014.00457

Friston, K. (2012). A free energy principle for biological systems. *Entropy* 14, 2100–2121. doi: 10.3390/e14112100

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O. Doherty, J., and Pezzulo, G. (2016). Active inference and learning. *Neurosci. Biobehav. Rev.* 68, 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput*. 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., Parr, T., and Zeidman, P. (2018). Bayesian model reduction. *arXiv [Preprint]* arXiv:1805.07092.

Friston, K., and Penny, W. (2011). Post hoc Bayesian model selection. *Neuroimage* 56, 2089–2099. doi: 10.1016/j.neuroimage.2011.03.062

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–224. doi: 10.1080/17588928.2015.1020053

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Front. Hum. Neurosci.* 7, 1–18. doi: 10.3389/fnhum.2013.00598

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017b). Active inference, curiosity and insight. *Neural Comput*. 29, 2633–2683. doi: 10.1162/neco_a_00999

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017a). Deep temporal models and active inference. *Neurosci. Biobehav. Rev*. 77, 388–402. doi: 10.1016/j.neubiorev.2017.04.009

Furuya, S., and Kinoshita, H. (2008). Expertise-dependent modulation of muscular and non-muscular torques in multi-joint arm movements during piano keystroke. *Neuroscience* 156, 390–402. doi: 10.1016/j.neuroscience.2008.07.028

Futuyma, D. J., and Moreno, G. (1988). The evolution of ecological specialization. *Annu. Rev. Ecol. Syst*. 19, 207–233. doi: 10.1146/annurev.es.19.110188.001231

Gentner, D. R. (1998). "Chapter 1: Expertise in typewriting," in *The Nature of Expertise*, eds M. T. H. Chi, R. Glaser, and M. J. Farr (Taylor & Francis Group), 1–21. doi: 10.4324/9781315799681

Gershman, S. J., and Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Curr. Opin. Neurobiol*. 20, 251–256. doi: 10.1016/j.conb.2010.02.008

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci*. 371:20110553. doi: 10.1098/rsta.2011.0553

Giedd, J. N. (2008). The teen brain: insights from neuroimaging. *J. Adolesc. Health* 42, 335–343. doi: 10.1016/j.jadohealth.2008.01.007

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66, 585–595. doi: 10.1016/j.neuron.2010.04.016

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.* doi: 10.1146/annurev.neuro.29.051605.112851

Harrold, L. R., Field, T. S., and Gurwitz, J. H. (1999). Knowledge, patterns of care, and outcomes of care for generalists and specialists. *J. Gen. Intern. Med*. 14, 499–511. doi: 10.1046/j.1525-1497.1999.08168.x

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Hobson, J. A., and Friston, K. J. (2012). Waking and dreaming consciousness: neurobiological and functional considerations. *Prog. Neurobiol*. 98, 82–98. doi: 10.1016/j.pneurobio.2012.05.003

Hoeting, J., Madigan, D., Raftery, A., and Volunsky, C. (1999). Bayesian model averaging: a tutorial. *Stat. Sci.* 14, 382–401.

Huttenlocher, P. R., and Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *J. Comp. Neurol*. 387, 167–178. doi: 10.1002/SICI1096-986119971020387:2<167::AID-CNE1>3.0.CO;2-Z

Huttenlocher, P. R., de Courten, C., Garey, L. J., and Van der Loos, H. (1982). Synaptogenesis in human visual cortex–evidence for synapse elimination during normal development. *Neurosci. Lett*. 33, 247–252. doi: 10.1016/0304-3940(82)90379-2

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artif. Intell*. 101, 99–134. doi: 10.1016/S0004-3702(98)00023-X

Kaplan, R., and Friston, K. J. (2018). Planning and navigation as active inference. *Biol. Cybern*. 112, 323–343. doi: 10.1007/s00422-018-0753-2

Keramati, M., Dezfouli, A., and Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput. Biol*. 7:e1002055. doi: 10.1371/journal.pcbi.1002055

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface* 15:20170792. doi: 10.1098/rsif.2017.0792

Klapp, S. T. (1995). Motor response programming during simple choice reaction time: the role of practice. *J. Exp. Psychol. Hum. Percept. Perform*. 21, 1015–1027. doi: 10.1037/0096-1523.21.5.1015

Krampe, R. T. (2002). Aging, expertise and fine motor movement. *Neurosci. Biobehav. Rev*. 26, 769–776. doi: 10.1016/S0149-7634(02)00064-7

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM J. Res. Dev*. 5, 183–191. doi: 10.1147/rd.53.0183

Linson, A., Clark, A., Ramamoorthy, S., and Friston, K. (2018). The active inference approach to ecological perception: general information dynamics for natural and artificial embodied cognition. *Front. Robot A. I*. 5:21. doi: 10.3389/frobt.2018.00021

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Front. Comput. Neurosci*. 10:56. doi: 10.3389/fncom.2016.00056

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*. 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Parr, T., and Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Sci. Rep*. 7, 1–21. doi: 10.1038/s41598-017-15249-0

Parr, T., and Friston, K. J. (2018). The computational anatomy of visual neglect. *Cereb. Cortex*. 28, 777–790. doi: 10.1093/cercor/bhx316

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–163.

Schwartenbeck, P., and Friston, K. (2016). Computational phenotyping in psychiatry: a worked example. *eNeuro* 3, 1–18. doi: 10.1523/ENEURO.0049-16.2016

Shiffrin, R. M., and Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychol. Rev.* 84, 127–190. doi: 10.1037/0033-295X.84.2.127

Sternberg, R. J., and Frensch, P. A. (1992). "On being an expert: a cost-benefit analysis," in *The Psychology of Expertise* (New York, NY: Springer New York), 191–203.

Tononi, G., and Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep Med. Rev*. 10, 49–62. doi: 10.1016/j.smrv.2005.05.002

Van Tienderen, P. H. (1991). Evolution of generalists and specialists in spatially heterogeneous environments. *Evolution* 45, 1317–1331. doi: 10.1111/j.1558-5646.1991.tb02638.x

Wu, A. W., Young, Y., Skinner, E. A., Diette, G. B., Huber, M., Peres, A., et al. (2001). Quality of care and outcomes of adults with asthma treated by specialists and generalists in managed care. *Arch. Intern. Med*. 161, 2554–2560. doi: 10.1001/archinte.161.21.2554

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 28 October 2020
doi: 10.3389/frai.2020.509354

Deep Active Inference and Scene Construction

R. Conor Heins[1,2,3], M. Berk Mirza[4,5,6], Thomas Parr[6], Karl Friston[6], Igor Kagan[3,7] and Arezoo Pooresmaeili[2,3]*

[1] Department of Collective Behaviour, Max Planck Institute for Animal Behavior, Konstanz, Germany, [2] Perception and Cognition Group, European Neuroscience Institute, A Joint Initiative of the University Medical Centre Göttingen and the Max-Planck-Society, Göttingen, Germany, [3] Leibniz Science Campus "Primate Cognition", Göttingen, Germany, [4] Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, [5] The National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre (BRC) at South London and Maudsley National Health Service (NHS) Foundation Trust and The Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, [6] Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom, [7] Decision and Awareness Group, Cognitive Neuroscience Laboratory, German Primate Centre (DPZ), Göttingen, Germany

Adaptive agents must act in intrinsically uncertain environments with complex latent structure. Here, we elaborate a model of visual foraging—in a hierarchical context—wherein agents infer a higher-order visual pattern (a "scene") by sequentially sampling ambiguous cues. Inspired by previous models of scene construction—that cast perception and action as consequences of approximate Bayesian inference—we use active inference to simulate decisions of agents categorizing a scene in a hierarchically-structured setting. Under active inference, agents develop probabilistic beliefs about their environment, while actively sampling it to maximize the evidence for their internal generative model. This approximate evidence maximization (i.e., self-evidencing) comprises drives to both maximize rewards and resolve uncertainty about hidden states. This is realized via minimization of a free energy functional of posterior beliefs about both the world as well as the actions used to sample or perturb it, corresponding to perception and action, respectively. We show that active inference, in the context of hierarchical scene construction, gives rise to many empirical evidence accumulation phenomena, such as noise-sensitive reaction times and epistemic saccades. We explain these behaviors in terms of the principled drives that constitute the *expected free energy*, the key quantity for evaluating policies under active inference. In addition, we report novel behaviors exhibited by these active inference agents that furnish new predictions for research on evidence accumulation and perceptual decision-making. We discuss the implications of this hierarchical active inference scheme for tasks that require planned sequences of information-gathering actions to infer compositional latent structure (such as visual scene construction and sentence comprehension). This work sets the stage for future experiments to investigate active inference in relation to other formulations of evidence accumulation (e.g., drift-diffusion models) in tasks that require planning in uncertain environments with higher-order structure.

Keywords: active inference, visual foraging, Bayesian brain, hierarchical inference, free energy, epistemic value, random dot motion

OPEN ACCESS

Edited by:
Nicole C. Kleinstreuer, National Institute of Environmental Health Sciences (NIEHS), United States

Reviewed by:
Guohua Huang, Shaoyang University, China
Dimitri Ognibene, University of Essex, United Kingdom

****Correspondence:***
R. Conor Heins
conor.heins@gmail.com

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *01 November 2019*
Accepted: *10 September 2020*
Published: *28 October 2020*

Citation:
Heins RC, Mirza MB, Parr T, Friston K, Kagan I and Pooresmaeili A (2020) Deep Active Inference and Scene Construction. Front. Artif. Intell. 3:509354. doi: 10.3389/frai.2020.509354

1. INTRODUCTION

Our daily life is full of complex sensory scenarios that can be described as examples of "scene construction" (Hassabis and Maguire, 2007; Zeidman et al., 2015; Mirza et al., 2016). In its most abstract sense, scene construction describes the act of inferring a latent variable (or "scene") given a set of (potentially ambiguous) sensory cues. Sentence comprehension is a prime example of scene construction: individual words are inspected in isolation, but after reading a sequence one is able to abduce the overall meaning of the sentence that the words are embedded within (Tanenhaus et al., 1995; Narayanan and Jurafsky, 1998; Ferro et al., 2010). This can be cast as a form of hierarchical inference in which low-level evidence (e.g., words) is actively accumulated over time to support disambiguation of high-level hypotheses (e.g., possible sentence meanings).

We investigate hierarchical belief-updating by modeling visual foraging as a form of scene construction, where individual images are actively sampled with saccadic eye movements in order to accumulate information and categorize the scene accurately (Yarbus, 1967; Jóhannesson et al., 2016; Mirza et al., 2016; Yang et al., 2016; Ólafsdóttir et al., 2019). In the context of scene construction, sensory uncertainty (e.g., blurry images) can limit the ability of individual cues to support inference about the overarching visual scene. Such sensory uncertainty can be partially "overridden" using prior knowledge, which might be built into the agent's internal model, innately or based on previous experience. While there is an enormous body of literature on the resolution of uncertainty with prior information (Trueswell et al., 1994; Rayner and Well, 1996; Körding and Wolpert, 2004; Stocker and Simoncelli, 2006; Girshick et al., 2011), relatively little research has examined interactions between sensory uncertainty and prior information in the context of a dynamic, active vision task like visual foraging or scene construction (with notable exceptions: e.g., Quétard et al., 2016).

Building on a previous Bayesian formulation of scene construction, in this work use we use *active inference* to model visual foraging in a hierarchical scene construction task (Friston et al., 2012a,b, 2017a; Mirza et al., 2016), and to study different types of uncertainty across distinct "layers of inference." We present simulations of active inference agents performing hierarchical scene construction while parametrically manipulating sensory uncertainty and prior beliefs. The (sometimes counterintuitive) results of our simulations invite new perspectives on active sensing and hierarchical inference, which we discuss in the context of experimental design for both visual foraging experiments and perceptual decision-making tasks more generally. We examine the model's behavior in terms of the tension between instrumental (or utility-driven) and exploratory (epistemically-driven) drives, and how active inference explains both by appealing to a single pseudo- "value function": the *expected free energy*.

The rest of this paper is structured as follows: first, we summarize active inference and the free energy principle, highlighting the *expected free energy*, a quantity that prescribes behavior with both goal-satisfying and information-gathering components, under the single theoretical mantle of maximizing model evidence. Next, we discuss the original model of scene construction that inspired the present work, and move on to introduce random dot motion stimuli and the ensuing ability to parametrically manipulate uncertainty across hierarchical levels, which distinguishes the current model from the original. Then we detail the Markov Decision Process generative model that our active inference agents entertain, and describe the belief-updating procedures used to invert generative models, given observed sensory data. Having appropriately set up our scene construction task, we then report the results of simulations, with differential effects of sensory uncertainty and prior belief strength appearing in several aspects of active evidence accumulation in this hierarchical environment. These computational demonstrations motivate our conclusion, where we discuss the implications of this work for experimental and theoretical studies of active sensing and evidence accumulation under uncertainty.

2. FREE ENERGY MINIMIZATION AND ACTIVE INFERENCE

2.1. Approximate Inference via Variational Bayes

The goal of Bayesian inference is infer possible explanations for data—this means obtaining a distribution over a set of parameters x (the causal variables or explanations), given some observations $\tilde{o}$, where the tilde $\sim$ notation indicates a sequence of such observations over time $\tilde{o} = [o_1, o_2, \ldots o_T]^T$. Note we use the notation x to refer to a set of causal variables, which may include (sequences of) states $\tilde{s}$ and/or hyperparameters. This is also called calculating the posterior probability $P(x|\tilde{o})$; it encodes the *optimal* belief about causal variables x, after having observed some data $\tilde{o}$. To compute the posterior requires solving using Bayes rule:

$$P(x|\tilde{o}) = \frac{P(\tilde{o}|x)P(x)}{P(\tilde{o})} \tag{1}$$

Importantly, computing this quantity requires calculating the marginal probability $P(\tilde{o})$, also known as the *evidence*:

$$P(\tilde{o}) = \sum_x P(x)P(\tilde{o}|x) \tag{2}$$

Solving this summation[1] (in the continuous case, integration) quickly becomes intractable for high-dimensional models, since the evidence needs to be calculated for every possible combination of parameters x. The marginalization in Equation (2) renders exact Bayesian inference expensive or impossible in many cases, motivating approximate inference methods. One of the leading classes of methods for approximate inference are the variational methods (Beal, 2004; Blei et al., 2017). Variational inference circumvents the issue of exact inference by introducing an arbitrary distribution $Q(x)$ to replace the true posterior. This replacement is often referred to as the *variational* or

[1]From now on we assume the use of discrete probability distributions for convenience and compatibility with the sort of generative models relevant to the current work.

approximate posterior. By constraining the form of the variational distribution, tractable schemes exist to optimize it in a way that (approximately) maximizes evidence. This optimization occurs with respect to a quantity called the *variational free energy*, which is a computable upper-bound on *surprise*, or the negative (log) evidence $-\ln P(\tilde{o})$. The relationship between surprise and free energy can be shown as follows using Jensen's inequality:

$$\begin{aligned} -\ln P(\tilde{o}) &= -\ln \sum_x P(\tilde{o}, x) \\ &= -\ln \sum_x Q(x) \frac{P(\tilde{o}, x)}{Q(x)} \\ &\leq -\sum_x Q(x) \ln \frac{P(\tilde{o}, x)}{Q(x)} = F \\ \implies -\ln P(\tilde{o}) &\leq F \end{aligned} \tag{3}$$

where F is the variational free energy and $P(\tilde{o}, x)$ is the joint probability of observations and hidden causes, also known as the *generative model*. The free energy can itself be decomposed into:

$$F = \mathrm{D}_{KL}[Q(x) \| P(x|\tilde{o})] - \ln P(\tilde{o}) \tag{4}$$

This decomposition allows us to see that the free energy becomes a tighter upper-bound on surprise the closer the variational distribution $Q(x)$ comes to the true posterior $P(x|\tilde{o})$, as measured by the Kullback-Leibler divergence[2]. When $Q(x) = P(x|\tilde{o})$, the divergence disappears and the free energy equals the negative log evidence, rendering inference exact. Variational inference is thus often described as the conversion of an integration problem (computing the marginal likelihood of observations as in Equation (2)) into an optimization problem, wherein the parameters of the variational distribution are changed to minimize F:

$$Q(x) = \arg\min_{Q(x)} F \approx P(x|\tilde{o}) \tag{5}$$

2.2. Active Inference and Expected Free Energy

Having discussed the variational approximation to Bayesian inference via free energy minimization, we now turn our attention to active inference. Active inference is a framework for modeling and understanding adaptive agents, premised on the idea that agents engage in approximate Bayesian inference with respect to an internal generative model of sensory data. Crucially, under active inference both action and perception are realizations of the single drive to minimize surprise. By using variational Bayesian inference to achieve this, an active inference agent generates Bayes-optimal beliefs about sources of variation in its environment by free-energy-driven optimization of an approximate posterior $Q(x)$. This can be analogized to the idea of perception as inference, wherein perception constitutes optimizing the parameters of an approximate posterior distribution over hidden states $Q(\tilde{s}|\pi)$, under a particular policy π[3]. In the context of neural systems, it is theorized that the parameters of these posterior beliefs about states are encoded by distributed neural activity in the agent's brain (Friston, 2008; Friston and Kiebel, 2009; Huang and Rao, 2011; Bastos et al., 2012; Parr and Friston, 2018c). Parameters of the generative model itself (such as the likelihood mapping $P(o|s)$) are hypothesized to be encoded by the network architectures, synaptic strengths, and neuromodulatory elements of the nervous system (Bogacz, 2017; Parr et al., 2018, 2019).

Action can also be framed as a consequence of variational Bayesian inference. Under active inference, policies (sequences of actions) correspond to sequences of "control states" —a type of hidden state that agents can directly influence. Actions are treated as samples from posterior beliefs about policies (Friston et al., 2012b). However, optimizing beliefs about policies introduces an additional complication. Optimal beliefs about hidden states $Q(\tilde{s})$ are a function of current and past observations. However, as the instantaneous free energy is a direct function of observations, it is not immediately clear how to optimize beliefs about policies when observations from the future are not available. This motivates the introduction of the *expected free energy*, or beliefs about the free energy expected in the future when pursuing a policy π. The free energy expected at future time point τ under a policy π is given by $G(\pi, \tau)$. Replacing the expectation over hidden states and outcomes in Equation (3) with the expectation over hidden states and outcomes in the future, we have:

$$G(\pi, \tau) = \sum_{o,s} Q(o_\tau, s_\tau|\pi) \ln \frac{Q(s_\tau|\pi)}{P(o_\tau, s_\tau)} \tag{6}$$

Here, we equip the agent with the prior belief that its policies minimize the free energy expected (under their pursuit) in the future. Under Markovian assumptions on the dependence between subsequent time points in the generative model $P(\tilde{o}, \tilde{s}|\pi) = \prod_t^T P(o_t|s_t)P(s_t|s_{t-1}, \pi)$ and a mean-field factorization of the approximate posterior across time (such that $Q(\tilde{s}|\pi) = Q(\pi)\prod_{\tau=1}^T Q(s_\tau)$), we can write the prior probability of a policy as proportional to the sum of the expected free energies over time under each policy:

$$P(\pi) \propto \exp(-\sum_\tau G(\pi, \tau)) \tag{7}$$

We will not derive the self-consistency of the prior belief that agents (believe they) will choose free-energy-minimizing policies, nor the full derivation of the expected free energy here. Interested readers can find the full derivations in Friston et al. (2015, 2017a) and Parr and Friston (2019). However, it is worth emphasizing that different components of the expected free energy clarify its implications for optimal behavior in active inference agents. These components are formally related to other discussions of adaptive behavior, such as the trade-off between exploration and

[2]The Kullback-Leibler divergence or *relative entropy* is a non-negative measure of dissimilarity between probability distributions.

[3]Hereafter we refer to observations and *hidden states* as o and s, respectively. We use the more generic term *hidden causes* x to refer to *all* aspects of the posterior—including hidden states, policies, and hyperparameters of the generative model.

exploitation. We can re-write the expected free energy for a given time-point τ and policy π as a bound on the sum of two expectations:

$$\begin{aligned}\mathbf{G}(\tau,\pi) &= \mathbb{E}_{Q(o_\tau,s_\tau|\pi)}[\ln Q(s_\tau|\pi) - \ln P(o_\tau,s_\tau)] \\ &\geq -\mathbb{E}_{Q(o_\tau|\pi)}[\mathrm{D}_{KL}[Q(s_\tau|o_\tau,\pi)||Q(s_\tau|\pi)]] \\ &\quad - \mathbb{E}_{Q(o_\tau|\pi)}[\ln P(o_\tau)] \end{aligned} \tag{8}$$

From this decomposition of the quantity bounded by the expected free energy G we illustrate the different kinds of "value" that contribute to behavior in active inference (Friston et al., 2013, 2015; Parr and Friston, 2017; Mirza et al., 2018). See the **Appendix** for a derivation of Equation (8). The left term on the RHS of the second line is a term that has been called *negative information gain*. Under active inference, the most likely policies are those that *minimize* the expected free energy of their sensory consequences—therefore, minimizing this left term promotes policies that disclose information about the environment by reducing uncertainty about the causes of observations, i.e., maximizing information gain. The right term on the RHS of the second line is often called negative extrinsic (or instrumental) value, and minimizing this term promotes policies that lead to observations that match the agent's prior expectations about observations, The relationship of these prior expectations to goal-directed behavior will become clear later in this section. We also offer an alternative decomposition of the expected free energy, formulating it in terms of minimizing a combination of *ambiguity* and *risk*:

$$\begin{aligned}\mathbf{G}(\tau,\pi) \geq & -\underbrace{\mathrm{E}_{Q(o_\tau|\pi)}[\mathrm{D}_{KL}[Q(s_\tau|o_\tau,\pi)\|Q(s_\tau|\pi)]]}_{\textit{Epistemic value}} \\ & - \underbrace{\mathrm{E}_{Q(o_\tau|\pi)}[\ln P(o_\tau)]]}_{\textit{Instrumental value}} \\ = & \underbrace{\mathbb{E}_{Q(s_\tau|\pi)}[\mathrm{H}[P(o_\tau|s_\tau)]]}_{\textit{Ambiguity}} + \underbrace{D_{KL}[Q(o_\tau|\pi)||P(o_\tau)]}_{\textit{Risk}} \end{aligned} \tag{9}$$

See the **Appendix** for a derivation of Equation (9). The first term on the RHS of the first line (previously referred to as information gain) we hereafter refer to as "epistemic value" (Friston et al., 2015; Mirza et al., 2016). It is equivalent to expected Bayesian surprise in other accounts of information-seeking behavior and curiosity (Linsker, 1990; Itti and Baldi, 2009; Gottlieb and Oudeyer, 2018). Such an epistemic drive has the effect of promoting actions that uncover information about hidden states via sampling informative observations. This intrinsic drive to uncover information, and its natural emergence via the minimization of expected free energy, is integral to accounts of exploratory behavior, curiosity, salience, and related active-sensing phenomena under active inference (FitzGerald et al., 2015; Friston et al., 2017b,d; Parr and Friston, 2017, 2018b; Mirza et al., 2019b). An alternative formulation of the expected free energy is given in the second line of Equation (9), where minimizing expected free energy promotes policies that reduce "ambiguity," defined as the expected uncertainty of observations, given the states expected under a policy. These notions of information gain and expected uncertainty will serve as a useful construct in understanding the behavior of active inference agents performing hierarchical scene construction later.

In order to understand how minimizing expected free energy **G** relates to the pursuit of preference-related goals or drives, we now turn to the second term on the RHS of the first line of Equation (9). In order to enable instrumental or "non-epistemic" goals to drive action, we supplement the agent's generative model with an unconditional distribution over observations $P(o)$ (sometimes called $P(o|m)$, where m indicates conditioning on the generative model of the agent)—this also factors into the log joint probability distribution in the first line of Equation (8). By fixing certain outcomes to have high (or low) probabilities as prior beliefs, minimizing **G** imbues action selection with an apparent instrumental or exploitative component, measured by how closely observations expected under a policy align with baseline expectations. Said differently: active inference agents pursue policies that result in outcomes that they *a priori* expect to encounter. The distribution $P(o)$ is therefore also often called the "prior preferences." Encoding preferences or desires as beliefs about future sensory outcomes underwrites the known duality between inference and optimal control (Todorov, 2008; Friston et al., 2009; Friston, 2011; Millidge et al., 2020). In the language of Expected Utility Theory (which explains behavior by appealing to the principle of maximizing expected rewards), the logarithm of such prior beliefs is equivalent to the utility function (Zeki et al., 2004). This component of **G** has variously been referred to as utility, extrinsic value, or instrumental value (Seth, 2015; Friston et al., 2017a; Biehl et al., 2018; Seth and Tsakiris, 2018); hereafter we will use the term instrumental value. A related but subtly different perspective is provided by the second term on the RHS of the second line of Equation (9): in this formulation, prior preferences enter the free energy through a "risk" term. The minimization of expected risk favors actions that minimize the KL-divergence between outcomes expected under a policy and preferred outcomes, and is related to formulations like KL-control or risk-sensitive control (Klyubin et al., 2005; van den Broek et al., 2010).

3. SCENE CONSTRUCTION WITH RANDOM DOT MOTION

3.1. The Original Model

We now describe an abstract scene construction task that will serve as the experimental context within which to frame our hierarchical account of active evidence accumulation. Inspired by a previous active inference model of scene construction introduced by Mirza et al. (2016), here we invoke scene construction in service of a categorization game. In each trial of the task, the agent must make a discrete choice to report its belief about the identity of the "hidden scene." In the formulation by Mirza et al., the scenes are represented by three abstract semantic labels: "flee," "feed," and "wait" (see **Figure 1**). Each scene manifests as a particular spatial coincidence of two pictures, where each picture is found within a single quadrant in a 2×2

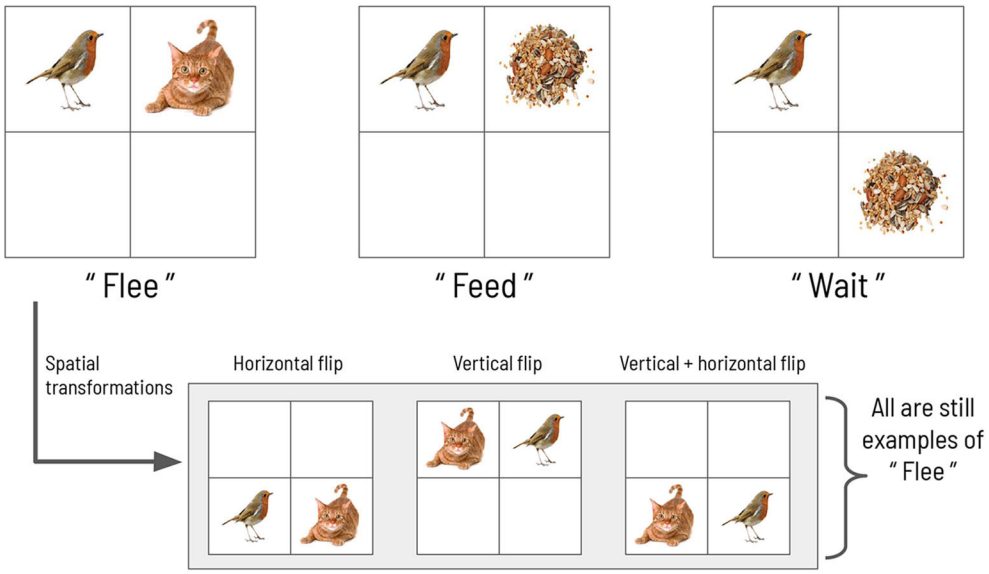


FIGURE 1 | The scene configurations of the original formulation. The three scenes characterizing each trial in the original scene construction study, adapted with permission from Mirza et al. (2016).

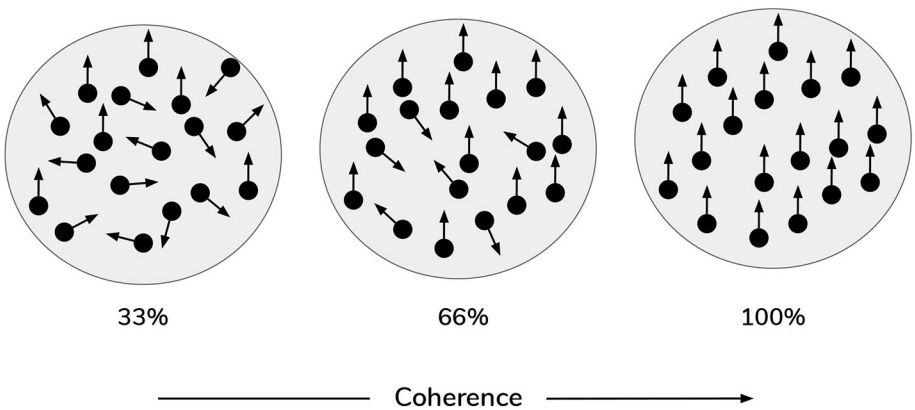


FIGURE 2 | Random Dot Motion Stimuli (RDMs). Schematic of random dot motion stimuli, with increasing coherence levels (i.e., % percentage of dots moving upwards) from left to right.

visual array. For example, the "flee" scene is defined as when a picture of a cat and a picture of a bird occupy two quadrants lying horizontally adjacent to one other. The scene identities are also invariant to two spatial transformations: vertical and horizontal inversions. For example, in the "flee" scene, the bird and cat pictures can be found in either in the top or bottom row of the 2×2 array, and they can swap positions; in any of these cases the scene is still "flee." The task requires active visual interrogation of the environment because quadrants must be *gaze-contingently* unveiled. That means, by default all quadrants are covered and their contents not visible; the agent must directly look at a quadrant in order to see its contents. This task structure and the ambiguous nature of the picture → scene mapping means that agents need to actively forage for information in the visual array in order to abduce the scene.

3.2. Introducing Random Dot Motion

In the current work, scene construction is also framed as a categorization task, requiring the gaze-contingent disclosure of quadrants whose contents furnish evidence for beliefs about the scene identity. However, in the new task, the visual stimuli occupying the quadrants are animated *random dot motion* or RDM patterns, instead of static pictographs. An RDM stimulus consists of a small patch of dots whose correlated displacement over time gives rise to the perception of apparent directed motion (see **Figure 2**). By manipulating the proportion of dots moving in the same direction, the apparent direction of motion can be made more or less difficult to discriminate (Shadlen and Newsome, 1996). This discriminability is usually operationalized as a single *coherence* parameter, which defines the percentage of dots that appear to move in a common direction. The

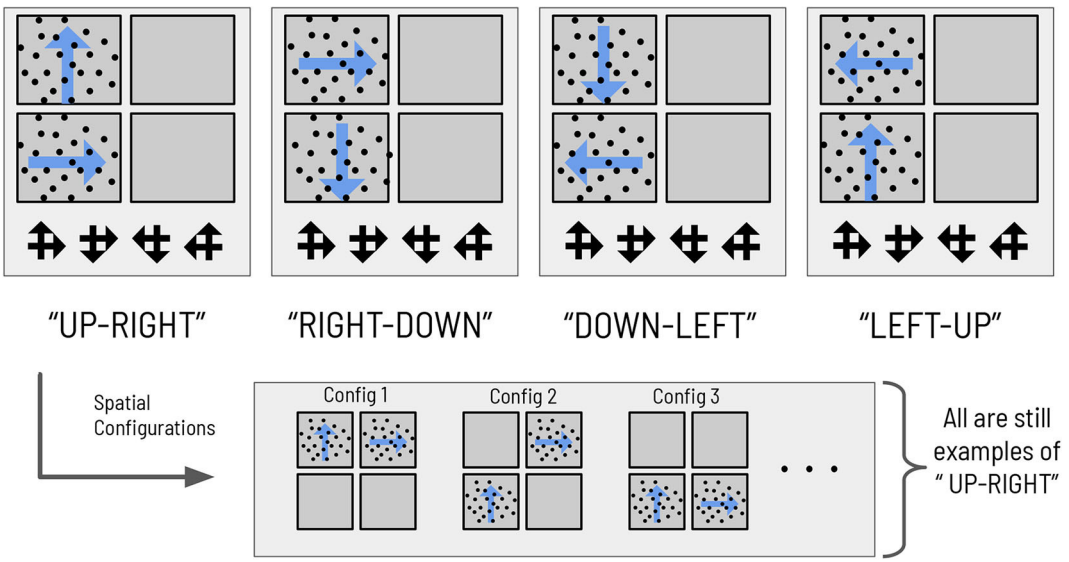


FIGURE 3 | The mapping between scenes and RDMs. The mapping between the four abstract scene categories and their respective dot motion pattern manifestations in the context of the hierarchical scene construction task. As an example of the spatial invariance of each scene, the bottom right panels show two possible (out of 12 total) RDM configurations for the scene "**RIGHT-DOWN**," where the two constitutive RDMs of that scene are found in exactly two of the four quadrants. The 'scene symbols' at the bottom of the visual array represent the categorization choices available to the subject, with each symbol comprised of two overlapping arrows that indicate the directions of the motion that define the scene.

remaining non-signal (or "incoherent") dots are usually designed to move in random independent directions. This coherence parameter thus becomes a simple proxy for sensory uncertainty in motion perception: manipulating the coherence of RDM patterns has well-documented effects on behavioral measures of performance, such as reaction time and discrimination accuracy, with increasing coherence leading to faster reaction times and higher accuracy (Palmer et al., 2005). In the current formulation, each RDM pattern is characterized by a unique primary direction of motion that belongs to one of the four cardinal directions: **UP**, **RIGHT**, **DOWN**, or **LEFT**. For example, in a given trial one quadrant may contain a motion pattern moving (on average) upwards, while another quadrant contains a motion pattern moving (on average) leftwards. These RDM stimuli are suitable for the current task because we can use the coherence parameter to tune motion ambiguity and hence sensory uncertainty. Applying this metaphor to the original task (Mirza et al., 2016): we might imagine blurred versions of the cat and bird pictures, such that it becomes difficult to tell whether a given image is of a bird or a cat—this low-level uncertainty about individual images may then "carry forward" to affect scene inference. An equivalent analogy might be found in the problem of reading a hastily-written phone number, such that it becomes hard to distinguish the number "7" from the number "1." In our case, the motion coherence of RDMs controls how easily an RDM of one direction can be confused with another direction—namely, a more incoherent dot pattern is more likely to be mistaken as a dot pattern moving in a different direction.

We also design the visual stimulus → scene mapping such that scenes are degenerate with respect to individual visual stimuli, as in the previous task (see **Figure 3**). There are four scenes, each one defined as the co-occurrence of two RDMs in two (and only two) quadrants of the visual array. The two RDMs defining a given scene move in perpendicular directions; the scenes are hence named: **UP-RIGHT**, **RIGHT-DOWN**, **DOWN-LEFT**, and **LEFT-UP**. Discerning the direction of one RDM is not sufficient to disambiguate the scene; due to the degeneracy of the scene configurations with respect to RDMs, the agent must always observe two RDMs and discern their respective directions before being able to unambiguously infer scene identity. The task requires two nested inferences—one about the contents of the currently-fixated quadrant (e.g., "Am I looking at an **UP**-wards moving RDM?") and another about the identity about the overarching scene (e.g., "is the scene **UP-RIGHT**?"). During each trial, an agent can report its guess about the scene identity by choosing one of the four symbols that signify the scenes (see **Figure 3**), which ends the trial. This concludes our narrative description of the experimental setup.

3.3. Summary

We have seen how both perception and action emerge as consequences of free energy minimization under active inference. Perception is analogized to state estimation and corresponds to optimizing variational beliefs about the hidden causes of sensory data x. Meanwhile actions are sampled from inferred sequences of control states (policies). The likelihood of a policy is inversely proportional to the free energy *expected* under that policy. We demonstrated that expected free energy can be decomposed into the sum of two terms, which respectively encode the drive to resolve ambiguity about the hidden causes of sensory data (epistemic value) and to satisfy agent-specific preferences (instrumental value) (first line of Equation 9). In

this way active inference theoretically dissolves the exploration-exploitation dilemma often discussed in decision sciences and reinforcement learning (March, 1991; Schmidhuber, 1991; Sutton and Barto, 1998; Parr, 2020) by choosing policies that minimize expected free energy. This unification of perception and action under a common Bayesian ontology underlies the power of active inference as a normative framework for studying adaptive behavior in complex systems. In the following sections we will present a (hierarchical) Markov Decision Process model of scene construction, where stochastic motion stimuli serve as observations for an overarching scene categorization task. We then discuss perception and action under active inference in the context of hierarchical scene construction, with accompanying computational demonstrations.

4. HIERARCHICAL MARKOV DECISION PROCESS FOR SCENE CONSTRUCTION

We now introduce the hierarchical active inference model of visual foraging and scene construction. The generative model (the agent) and the generative process of the environment both take the form of a Markov Decision Process or MDP. MDPs are a simple class of probabilistic generative models where space and time are treated discretely (Puterman, 1995). In the MDP used here, states are treated as discrete samples from categorical distributions and likelihoods act as linear transformations of hidden states, mapping states at one time step to the subsequent time step, i.e., $P(s_t|s_{t-1})$. This specification imbues the environment with Markovian, or "memoryless" dynamics. An extension of the standard MDP formulation is the *partially-observed* MDP or POMDP, which includes discrete observations that are mapped (via a likelihood function $P(o_t|s_t)$) from states to observations at a given time.

A generative model is simply a joint probability distribution over sensory observations and their latent causes $P(\tilde{o}, x)$, and is often factorized into the product of a likelihood and a set of marginal distributions over latent variables and hyperparameters, e.g., $P(\tilde{o}|\tilde{s})P(\tilde{s})P(\varphi)P(\zeta)\ldots$ where $\tilde{s}, \varphi, \zeta, \ldots \in x$ refer to the various latent causes. Note that in the current formulation the only hidden variables subject to variational inference are hidden states $\tilde{s}$ and policies π. The discrete MDP constrains these distributions to have a particular form; here, the priors over initial states, transition and likelihood matrices are encoded as categorical distributions over a discrete set of states and observations. Agents can only directly observe sensory outcomes $\tilde{o}$, meaning that the agent must infer hidden states $\tilde{s}$ by inverting the generative model to estimate the causes of observations. Hierarchical models take this a step further by adding multiple layers of hidden-state inference, allowing beliefs about hidden states $\tilde{s}^{(i)}$ at one level to act as so-called "inferred observations" $\tilde{o}^{(i+1)}$ for the level above, with associated priors and likelihoods operating at all levels. This marks a departure from previous work in the hierarchical POMDP literature (Pineau et al., 2001; Theocharous et al., 2004), where the hierarchical decomposition of *action* is emphasized and used to finesse the exponential costs of planning; states and observations, on the other hand, are often coarse-grained using separate schemes or left fully enumerated (although see Sridharan et al., 2010). In the current formulation, we adopt a hybrid scheme, where at a given level of depth in the hierarchy, observations can be *both* passed in at the same level (from the generative process), as well as via "inferred" observations from the level below. Note that as with $\tilde{o}$, we use $\tilde{s}$ to denote a sequence of hidden states over time $\tilde{s} = [s_1, s_2, \ldots s_T]^T$.

4.1. Hierarchical MDPs

Figure 4 summarizes the structure of a generic two-layer hierarchical POMDP model, outlining relationships between random variables via a Bayesian graph and their (factorized, categorical) forms in the left panel. In the left panel of **Figure 4**, $\tilde{o}$ and $\tilde{s}$ indicate sequences of observations and states over time. In the MDP model, the probability distributions that involve these sequences are expressed in a factorized fashion. The model's beliefs about how hidden states $\tilde{s}^{(i)}$ cause observations $\tilde{o}^{(i)}$ are expressed as multidimensional arrays in the likelihood matrix $\mathbf{A}^{(i),m}$, where i indicates the index of the hierarchical level and m indicates a particular modality (Mirza et al., 2016; Friston et al., 2017d). The (x, y) entry of a likelihood matrix $\mathbf{A}^{(i),m}$ prescribes the probability of observing the outcome x under the modality m at level i, given hidden state y. In this way, the columns of the $\mathbf{A}$ matrices are conditional categorical distributions over outcomes, given the hidden state indexed by the column. The dynamics that describe how hidden states at a given level $s^{(i)}$ evolve over time are given by Markov transition matrices $\mathbf{B}^{(i),n}(u)$ which express how likely the next state is given the current state—in the generative model, this is equivalent to the transition distribution $P(s_t|s_{t-1}, u_t)$. Here n indexes a particular factor of level i's hidden states, and u indexes a particular control state or action. Actions in this scheme are thus treated as controlled transitions between hidden states. We assume that the posterior distribution over different dimensions of hidden states factorize, leading to conditional independence between separate hidden state factors. This is known as the *mean-field approximation*, and allows the sufficient statistics of posterior beliefs about different hidden state variables to be updated separately (Feynman, 1998). This results in a set of relatively simple update equations for posterior beliefs and is also consistent with known features of neuroanatomy, e.g., functional segregation in the brain (Felleman and Van, 1991; Ungerleider and Haxby, 1994; Friston and Buzsáki, 2016; Mirza et al., 2016; Parr and Friston, 2018a). The hierarchical MDP formulation notably permits a segregation of timescales across layers and an according mean-field approximation on their respective free energies, such that multiple time steps of belief-updating at one level can unfold within a single time step of inference at the level above. In this way, low-level beliefs about hidden states (and policies) can be accumulated over time at a lower layer, at which point the final posterior estimate about hidden states is passed "up" as an *inferred* outcome to the layer above. Subsequent layers proceed at their own characteristic (slower) timescales (Friston et al., 2017d) to update beliefs about hidden states at their respective levels. Before we describe the particular form of the hierarchical MDP used (as both the generative process and generative model) for deep scene

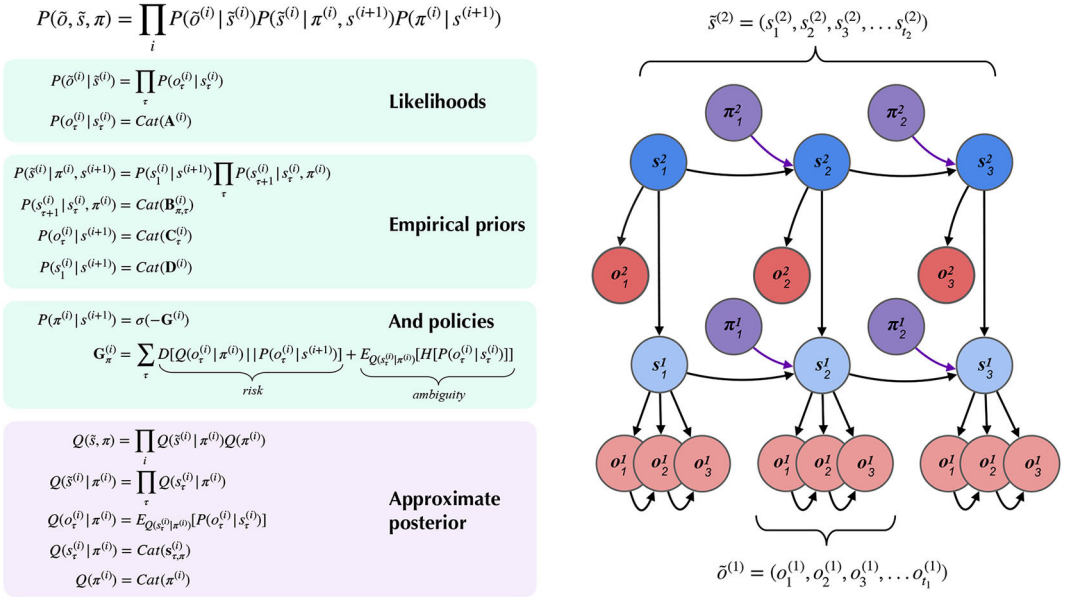


FIGURE 4 | A partially-observed Markov Decision Process with two hierarchical layers. Schematic overview of the generative model for a hierarchical partially-observed Markov Decision Process. The generic forms of the likelihoods, priors, and posteriors at hierarchical levels are provided in the left panels, adapted with permission from Friston et al. (2017d). *Cat*(**x**) indicates a categorical distribution, and $\tilde{x}$ indicates a discrete sequence of states or random variables: $\tilde{x} = (x_1, x_2, \ldots, x_t)$. Note that priors at the highest level (Level 2) are not shown, but are unconditional (non-empirical) priors, and their particular forms for the scene construction task are described in the text. As shown in the "Empirical Priors" panel, prior preferences at lower levels $\mathbf{C}_\tau^{(i)}$ can be a function of states at level $i+1$, but this conditioning of preferences is not necessary, and in the current work we pre-determine prior preferences at lower levels, i.e., they are not contextualized by states at higher levels (see **Figure 8**). Posterior beliefs about policies are given by a softmax function of the expected free energy of policies at a given level. The approximate (variational) beliefs over hidden states are represented via a mean-field approximation of the full posterior, such that hidden states can be encoded as the product of marginal distributions. Factorization of the posterior is assumed across hierarchical layers, across hidden state factors (see the text and **Figures 6**, **7** for details on the meanings of different factors), and across time. "Observations" at the higher level ($\tilde{o}^{(2)}$) may belong to one of two types: (1) observations that directly parameterize hidden states at the lower level via the composition of the observation likelihood one level $P(o^{(i+1)}|s^{(i+1)})$ with the empirical prior or "link function" $P(s^{(i)}|o^{(i+1)})$ at the level below, and (2) observations that are directly sampled *at the same level* from the generative process (and accompanying likelihood of the generative model $P(o^{(i+1)}|s^{(i+1)})$). For conciseness, we represent the first type of mapping, from states at $i+1$ to states at i through a direct dependency in the Bayesian graphical model in the right panel, but the reader should note that in practice this is achieved via the composition of two likelihoods: the observation likelihood at level $i+1$ and the link function at level i. This composition is represented by a single empirical prior $P(s^{(i)}|s^{(i+1)}) = Cat(\mathbf{D}^{(i)})$ in the left panel. In contrast, all observations at the lowest level ($\tilde{o}^{(1)}$) feed directly from the generative process to the agent.

construction, we provide a brief technical overview of the update scheme used to solve POMDPs with active inference.

4.1.1. Belief Updating

Figure 5 provides a schematic overview of the belief update equations for state estimation and policy inference under active inference. For the sake of clarity here we only consider a single "layer" of a POMDP generative model, i.e., we don't include the top-down or bottom-up beliefs that parameterize priors over hidden states (from the layer above) or inferred observations (from the layer below). Note that in this formulation, instead of directly evaluating the solution for states with lowest free energy $\mathbf{s}^*$, we use a marginal message passing routine to perform a gradient descent on the variational free energy at each time step, where posterior beliefs about hidden states and policies are incremented using prediction errors ε (see **Figure 5** legend for more details). In the context of deep temporal models, these equations proceed independently at each level of the hierarchy at each time step. At lower levels, the posterior over certain hidden state factors at the first timestep $\mathbf{s}_1^{(i)}$ can be initialized as the "expected observations" $\mathbf{o}^{(i+1)}$ from the level above, and "inferred observations" at higher levels are inherited as the final posterior beliefs $\mathbf{s}_T^{(i)}$ over the corresponding hidden state at lower levels. This update scheme may sound complicated; however, when expressed as a gradient descent on free energy, with respect to the sufficient statistics of beliefs about expected states, it reduces to a remarkably simple scheme that bears resemblance to neuronal processing: see Friston et al. (2015) for details. Importantly, the mean-field factorization of the generative model across hierarchical layers allows the belief updating to occur in isolation at each layer of the hierarchy, such that only the final posterior beliefs at one layer need to be passed to the layer above. The right side of **Figure 5** shows a simple schematic of how the particular random variables that make up generative model might correspond to neural processing in known brain regions. Evidence for the sort of hierarchical processing entailed by such generative models abounds in the brain, and is the subject of a wealth of empirical and theoretical neuroscience research (Lee and Mumford, 2003; Friston, 2008; Hasson et al., 2008; Friston et al., 2017c; Runyan et al., 2017; Pezzulo et al., 2018).

A Perception and state estimation

$$\mathbf{s}^* = \text{argmin} F(\pi) = \sigma(\ln \mathbf{B}_{\pi,\tau-1}\mathbf{s}_{\pi,\tau-1} + \ln \mathbf{B}_{\pi,\tau} \cdot \mathbf{s}_{\pi,\tau+1} + [\tau \leq t] \cdot \zeta \ln \mathbf{A} \cdot o_\tau)$$

$$\epsilon_\tau^\pi = \ln \mathbf{s}^* - \ln \mathbf{s}_\tau^\pi$$

$$\mathbf{s}_\tau^\pi = \sigma(\mathbf{v}_\tau^\pi) : \dot{\mathbf{v}}_\tau^\pi = \epsilon_\tau^\pi$$

B Policy evaluation and selection

$$\boldsymbol{\pi} = \sigma(-\mathbf{F} - \gamma \cdot \mathbf{G}) \qquad \boldsymbol{\pi} = Q(\pi)$$

$$\mathbf{F}_\pi = \sum_\tau \epsilon_\tau^\pi \cdot \mathbf{s}_\tau^\pi$$

$$\mathbf{G}_\pi = \sum_\tau (\mathbf{o}_\tau^\pi \cdot (\ln \mathbf{o}_\tau^\pi - \mathbf{C}_\tau) + \mathbf{H} \cdot \mathbf{s}_\tau^\pi)$$

C Action selection (and model averaging)

$$a_t = \max_a \boldsymbol{\pi} \qquad \boldsymbol{\pi}_\pi = (u_{\pi,1}, u_{\pi,2}, u_{\pi,3}, \ldots, u_{\pi,T})$$

$$\mathbf{s}_\tau = \sum_\pi \boldsymbol{\pi}_\pi \cdot \mathbf{s}_\tau^\pi$$

$$\mathbf{o}_\tau^\pi = \mathbf{A}\mathbf{s}_\tau^\pi$$

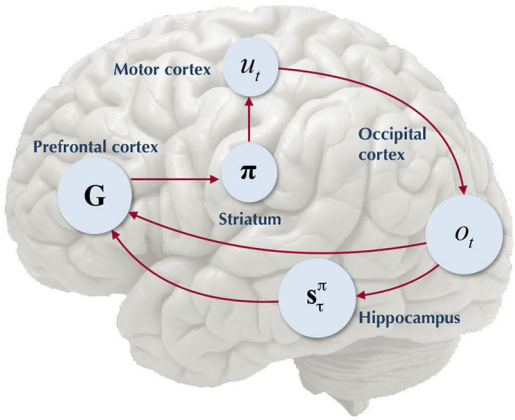


FIGURE 5 | Belief-updating under active inference. Overview of the update equations for posterior beliefs under active inference. **(A)** Shows the optimal solution for posterior beliefs about hidden states $\mathbf{s}^*$ that minimizes the variational free energy of observations. In practice the variational posterior over states is computed as a marginal message passing routine (Parr et al., 2019), where prediction errors ε_τ^π minimized over time until some criterion of convergence is reached ($\varepsilon \approx 0$). The prediction errors measure the difference between the current log posterior over states $\ln \mathbf{s}_\tau^\pi$ and the optimal solution $\ln \mathbf{s}^*$. Solving via error-minimization lends the scheme a degree of biological plausibility and is consistent with process theories of neural function like predictive coding (Bastos et al., 2012; Bogacz, 2017). An alternative scheme would be equating the marginal posterior over hidden states (for a given factor and/or timestep) to the optimal solution $\mathbf{s}_{\pi,\tau}^*$—this is achieved by solving for $\mathbf{s}^*$ when free energy is at its minimum (for a particular marginal), i.e., $\frac{\partial F}{\partial \mathbf{s}_{\pi,\tau}} = 0$. This corresponds to a fixed-point minimization scheme (also known as coordinate-ascent iteration), where each conditional marginal is iteratively fixed to its free-energy minimum, while holding the remaining marginals constant (Blei et al., 2017). **(B)** Shows how posterior beliefs about policies are a function of the free energy of states expected under policies **F** and the expected free energy of policies **G**. **F** is a function of state prediction errors and expected states, and **G** is the expected free energy of observations under policies, shown here decomposed into the KL divergence between expected and preferred observations or risk ($\mathbf{o}_\tau^\pi \cdot (\ln \mathbf{o}_\tau^\pi - \mathbf{C}_\tau)$) and the expected entropy or ambiguity ($\mathbf{H} \cdot \mathbf{s}_\tau^\pi$). A precision parameter γ scales the expected free energy and serves as an inverse temperature parameter for a softmax normalization σ of policies. See the text (Section 4.1.1) for more clarification on the free energy of policies **F**. **(C)** Shows how actions are sampled from the posterior over policies, and the posterior over states is updated via a Bayesian model average, where expected states are averaged under beliefs about policies. Finally, expected observations are computed by passing expected states through the likelihood of the generative model. The right side shows a plausible correspondence between several key variables in an MDP generative model and known neuroanatomy. For simplicity, a hierarchical generative model is not shown here, but one can easily imagine a hierarchy of state inference that characterizes the recurrent message passing between lower-level occipital areas (e.g., primary visual cortex) through higher level visual cortical areas, and terminating in "high-level," prospective and policy-conditioned state estimation in areas like the hippocampus. We note that it is an open empirical question, whether various computations required for active inference can be localized to different functional brain areas. This figure suggests a simple scheme that attributes different computations to segregated brain areas, based on their known function and neuroanatomy (e.g., computing the expected free energy of actions (G), speculated to occur in frontal areas).

We also find it worthwhile to clarify the distinction between the *variational free energy* of policies $F(\pi)$ and the *expected free energy* of policies $G(\pi)$, both of which are needed to compute the posterior over policies $Q(\pi)$. The final posterior probability over policies is a softmax function of both quantities (see **Figure 5B**), where the former can be seen as the evidence afforded by past and ongoing observations, that a given policy is *currently* being pursued, whereas the latter is the evidence *expected* to be gathered in favor of pursuing a given policy, where this expected evidence is biased by prior beliefs about what kinds of observations the agent is likely to encounter (via the prior preferences **C**). Starting with the definition of the free energy of the (approximate) posterior over both hidden states and policies::

$$F = \mathrm{E}_{Q(\tilde{s},\pi)}[\ln Q(\tilde{s},\pi) - \ln P(\tilde{o},\tilde{s},\pi)]$$

$$= \mathrm{E}_{Q(\pi)}[F(\pi)] + \mathrm{D}_{KL}[Q(\pi)||P(\pi)] \tag{10}$$

$$F(\pi) = -\mathrm{E}_{Q(\tilde{s}|\pi)}\left[\ln P(\tilde{o},\tilde{s}|\pi)\right] - \mathrm{H}[Q(\tilde{s}|\pi)]$$
$$Q(\pi) = \arg\min_{Q(\pi)} F \propto e^{(\ln P(\pi) - F(\pi))} \tag{11}$$

Where $\ln P(\pi) = G(\pi)$ is a prior of the generative model that encodes the self-consistent belief that the prior probability of a policy is proportional to its negative expected free energy $G(\pi)$. Please see the **Appendix** for a fuller derivation of Equation (10). Note that (due to the factorization of the approximate posterior over time, cf. Section 2.2) the variational free energy of a policy $F(\pi)$ is the sum of the individual free energies for a given policy afforded by past observations, up to and including the current observation:

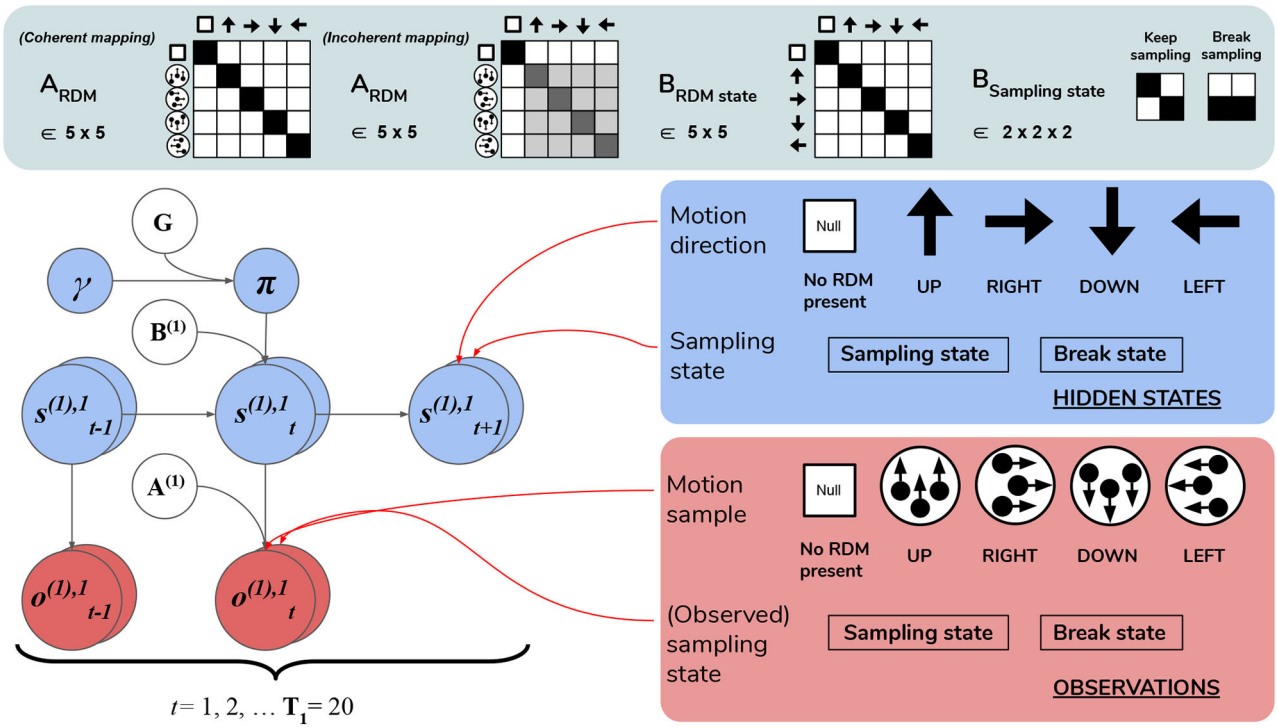


FIGURE 6 | Level 1 MDP. **Level 1** of the hierarchical POMDP for scene construction (see Section 4.2.1 for details). Level 1 directly interfaces with stochastic motion observations generated by the environment. At this level hidden states correspond to: (1) the true motion direction $\mathbf{s}^{(1),1}$ underlying visual observations at the currently-fixated region of the visual array and (2) the sampling state $\mathbf{s}^{(1),2}$, an aspect of the environment that can be changed via actions, i.e., selections of the appropriate state transition, as encoded in the **B** matrix. The first hidden state factor $\mathbf{s}^{(1),1}$ can either correspond to a state with no motion signal ("Null," in the case when there is no RDM or a categorization decision is being made) or assume one of the four discrete values corresponding to the four cardinal motion directions. At each time step of the generative process, the current state of the RDM stimulus $\mathbf{s}^{(1),1}$ is probabilistically mapped to a motion observation via the first-factor likelihood $\mathbf{A}^{(1),1}$ (shown in the top panel as $\mathbf{A}_{\text{RDM}}$). The entropy of the columns of this mapping can be used to parameterize the coherence of the RDM stimulus, such that the true motion states $\mathbf{s}^{(1),1}$ cause motion observations $\mathbf{o}^{(1),1}$ with varying degrees of fidelity. This is demonstrated by two exemplary $\mathbf{A}_{\text{RDM state}}$ matrices in the top panel (these correspond to $\mathbf{A}^{(1),1}$): the left-most matrix shows a noiseless, "coherent" mapping, analogized to the situation of when an RDM consists of all dots moving in the same direction as described by the true hidden state; the matrix to the right of the noiseless mapping corresponds to an incoherent RDM, where instantaneous motion observations may assume directions different than the true motion direction state, with the frequency of this deviation encoded by probabilities stored in the corresponding column of $\mathbf{A}_{\text{RDM}}$. The motion direction state doesn't change in the course of a trial (see the identity matrix shown in the top panel as $\mathbf{B}_{\text{RDM}}$, which simply maps the hidden state to itself at each subsequent time step)—this is true of both the generative model and the generative process. The second hidden state factor $\mathbf{s}^{(1),2}$ encodes the current "sampling state" of the agent; there are two levels under this factor: "**Keep-sampling**" or "**Break-sampling**." This sampling state (a factor of the generative process) is directly represented as a control state in the generative model; namely, the agent can change it by sampling actions (B-matrix transitions) from the posterior beliefs about policies. The agent believes that the "**Break-sampling**" state is a *sink* in the transition dynamics, such that once it is entered, it cannot be exited (see the right-most matrix of the transition likelihood $\mathbf{B}_{\text{Sampling state}}$). Entering the "**Break-sampling**" state terminates the POMDP at Level 1. The "**Keep-sampling**" state enables the continued generation of motion observations as samples from the likelihood mapping $\mathbf{A}^{(1),1}$. $\mathbf{A}^{(1),2}$ (the "proprioceptive" likelihood, not shown for clarity) deterministically maps the current sampling state $\mathbf{s}^{(1),2}$ to an observation $\mathbf{o}^{(1),2}$ thereof (bottom row of lower right panel), so that the agent always observes which sampling state it is in unambiguously.

$$
\begin{aligned}
F(\pi) &= \sum_{\tau} F(\pi, \tau) \\
F(\pi, \tau) &= -\mathrm{E}_{Q(s_\tau|\pi)Q(s_{\tau-1}|\pi)}\big[[\tau \leq t] \ln P(o_\tau|s_\tau) \\
&\quad + \ln P(s_\tau|s_{\tau-1}, \pi) - \ln Q(s_\tau|\pi)\big] \qquad (12)
\end{aligned}
$$

The Iverson brackets $[\tau \leq t]$ return 1 if $\tau \leq t$ and 0 otherwise.

4.2. From Motion Discrimination to Scene Construction: A Nested Inference Problem

We now introduce the deep, temporal model of scene construction using the task discussed in Section 3 as our example (**Figure 6**). We formulate perception and action with a hierarchical POMDP consisting of two distinct layers that are solved via active inference. The first, shallowest level (Level 1) is an MDP that updates posterior beliefs about the most likely cause of visual stimulation (RDM direction), where we model the ongoing contents of single fixations—the stationary periods of relative retinal-stability between saccadic eye movements. This inference is achieved with respect to the (spatially-local) visual stimuli underlying current foveal observations. A binary policy is also implemented, encoding the option to continue holding fixation (and thus keep sampling the current stimulus) or to interrupt sampling and terminate updates at the lower level. The second, higher level (Level 2) is another MDP that performs inference at a slower timescale, with respect to the overarching

hidden scene that describes the current trial. Here, we enable policies that realize visual foraging. These policies encode controlled transitions between different states of the oculomotor system, serving as a model of saccadic eye movements to different parts of the visual array. This method of encoding saccades as controlled transitions between locations is inspired by the original scene construction formulation in Mirza et al. (2016). We will now discuss both layers individually and translate different elements of the MDP generative model and environment to task-relevant parameters and the beliefs of the agent.

4.2.1. Level 1: Motion Discrimination via Motion Sampling Over Time

Lowest level (Level 1) beliefs are updated as the agent encounters a stream of ongoing, potentially ambiguous visual observations—the instantaneous contents of an individual fixation. The hidden states at this level describe a distribution over motion directions, which parameterize the true state of the random motion stimulus within the currently-fixated quadrant. Observations manifest as a sequence of stochastic motion signals that are samples from the true hidden state distribution.

The generative model has an identical form as the generative process (see above) used to generate the stream of Level 1 outcomes. Namely, it is comprised of a set of likelihoods and transitions as the dynamics describing the "real" environment (**Figure 6**). In order to generate a motion observation, we sample the probability distribution over motion direction given the true hidden state using the Level 1 generative process likelihood matrix $\mathbf{A}^{(1),1}$. For example, if the current true hidden state at the lower level is **2** (implying that an RDM stimulus of **UP**wards motion occupies the currently fixated quadrant), stochastic motion observations are sampled from the *second* column of the generative likelihood mapping $\mathbf{A}^{(1),1}$. The precision of this column-encoded distribution over motion observations determines how often the sampled motions will be **UP**wards signals and thus consistent with the true hidden state. The entropy or ambiguity of this likelihood mapping operationalizes sensory uncertainty and in this case, motion incoherence. For more details on how states and outcomes are discretized in the generative process, see **Figure 6** and its legend.

Inference about the motion direction (Level 1 state estimation) roughly proceeds as follows: (1) at time t a motion observation $o_t^{(1),1}$ is sampled from the generative process $\mathbf{A}^{(1),1}$; (2) posterior beliefs about the motion direction at the current timestep $s_t^{(1),1}$ are updated using a gradient descent on the variational free energy. In addition, we included a second, controllable hidden state factor at Level 1 that we refer to as the abstract "sampling state" of the agent. We include this in order to enable policies at this level, which entail transitions between the two possible values of this control state. These correspond to the choice to either keep sampling the current stimulus or break sampling. These policies are stored as two 2×2 transition matrices in $\mathbf{B}^{(2),2}$, where each transition matrix $\mathbf{B}^{(2),2}(u)$encodes the probability of transitioning to "**Keep-sampling**" or "**Break-sampling**," given an action u and occupancy in one of the two sampling states. Note that these policies only consider actions at the next time step, meaning that the policy-space is identical to the action-space (there is no sequential aspect to the policies). Selecting the first action keeps the Level 1 MDP in the "**Keep-sampling**" state, triggering the generation of another motion observation from the generative process. Engaging the second "**Break-sampling**" policy moves the agent's sampling regime into the second state and terminates any further updates at Level 1. At this point the latest posterior beliefs from Level 1 are sent up as observations for Level 2. It is worth noting that implementing "breaking" the MDP at the lower level as an explicit policy departs from the original formulation of deep, temporal active inference. In the formulation developed in Friston et al. (2017d), termination of lower level MDPs occurs once the entropy of the lower-level posterior over the hidden states (only those factors that are linked with the level above) is minimized beyond a fixed value[4]. We chose to treat breaking the first level MDP as an explicit policy in order to formulate behavior in terms of the same principles that drive action selection at the higher level—namely, the expected free energy of policies. In the Simulations section we explore how the dynamic competition between the "**Break-**" and "**Keep-sampling**" policies induces an unexpected distribution of break latencies.

We fixed the maximum temporal horizon of Level 1 (hereafter T_1) to be 20 time steps, such that if the "**Break-sampling**" policy is not engaged before $t = 20$ (implying that "**Keep-sampling**" has been selected the whole time), Level 1 automatically terminates after the 20$^{\text{th}}$ time step and the final posterior beliefs are passed up as outcomes for Level 2.

4.2.2. Level 2: Scene Inference and Saccade Selection

After beliefs about the state of the currently-foveated visual region are updated via active inference at Level 1, the resulting posterior belief about motion directions is passed up to Level 2 as a belief about observations. These observations (which can be thought of as the inferred state of the visual stimulus at the foveated area) are used to update the statistics of posterior beliefs over the hidden states operating at Level 2 (specifically, the hidden state factor that encodes the identity of the scene, e.g., **UP-RIGHT**). Hidden states at Level 2 are segregated into two factors, with corresponding posterior beliefs about them updated independently.

The first hidden state factor corresponds to the scene identity. As described in Section 3, there are four possible scenes characterizing a given trial: **UP-RIGHT**, **RIGHT-DOWN**, **DOWN-LEFT**, and **LEFT-UP**. The scene determines the identities of the two RDMs hiding throughout the four quadrants, e.g., when the scene is **UP-RIGHT**, one **UP**wards-moving RDM is found in one of the four quadrants, and a **RIGHT**wards-moving RDM is found in another quadrant. The quadrants that are occupied by RDMs for a given scene is random, meaning that agents have to forage the 2×2 array for the RDMs in order to infer the scene. We encode the scene identities as

[4]This threshold is referred to as "residual uncertainty," and by default is set to as $\frac{1}{64}$ nats.

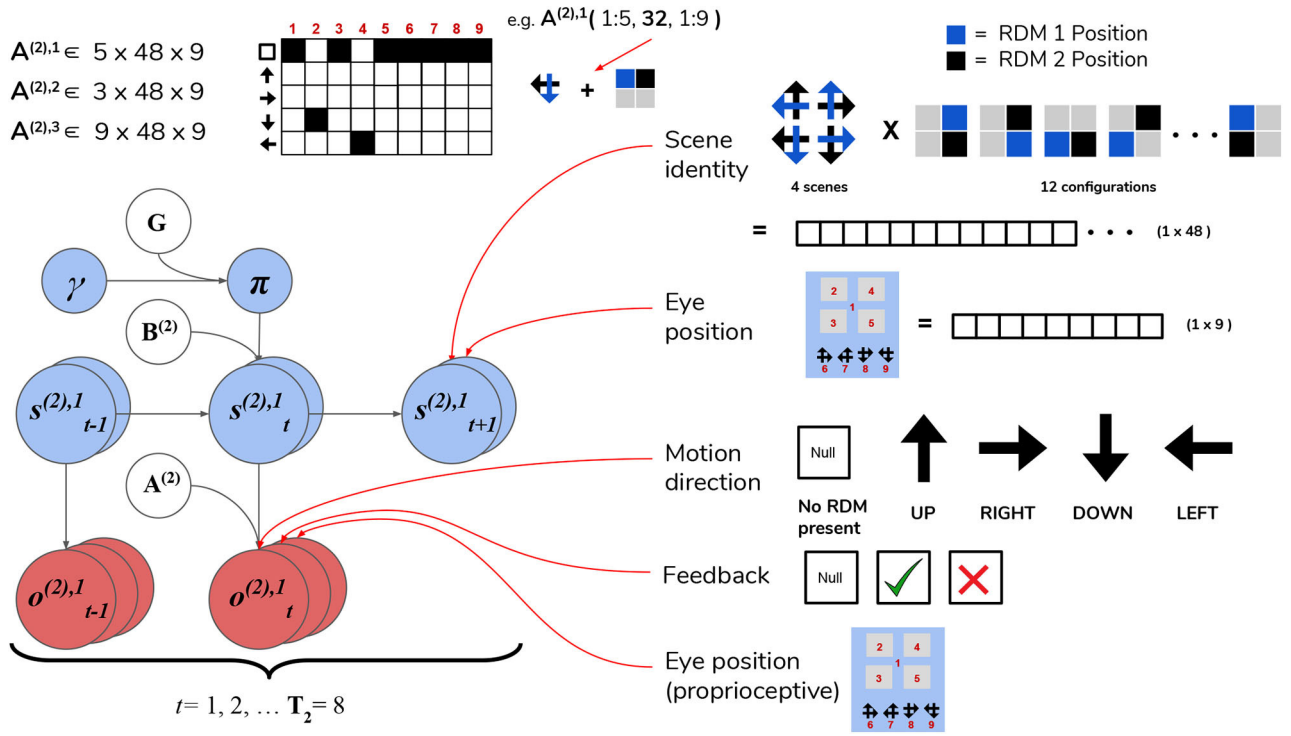


FIGURE 7 | Level 2 MDP. **Level 2** of the hierarchical POMDP for scene construction. Hidden states consist of two factors, one encoding the scene identity and another encoding the eye position (i.e., current state of the oculomotor system). The first hidden state factor $\mathbf{s}^{(2),1}$ encodes the scene identity of the trial in terms of two unique RDM directions occupy two of the quadrants (four possible scenes as described in the top right panel) and spatial configuration (one of 12 unique ways to place two RDMs in four quadrants). This yields a dimensionality of 48 for this hidden state factor (4 scenes × 12 spatial configurations). The second hidden state factor $\mathbf{s}^{(2),2}$ encodes the eye position, which is initialized to be in the center of the quadrants (Location 1). The next four values of this factor index the four quadrants (2–5), and the last four are indices for the choice locations (the agent fixates one of these four options to guess the scene identity). As with the sampling state factor at Level 1, the eye position factor $\mathbf{s}^{(2),2}$ is controllable by the agent through the action-dependent transition matrices $\mathbf{B}^{(2),2}$. Outcomes at Level 2 are characterized by three modalities: the first modality $\mathbf{o}^{(2),1}$ indicates the visual stimulus (or lack thereof) at the currently-fixated location. Note that during belief updating, the observations of this modality $\mathbf{o}^{(2),1}$ are inferred hidden states over motion directions that are passed up after solving the Level 1 MDP (see **Figure 6**). An example likelihood matrix for this first modality is shown in the upper left, showing the conditional probabilities for visual outcomes when the 1st factor hidden state has the value 32. This corresponds to the scene identity **DOWN-LEFT** under spatial configuration 8 (the RDMs occupy quadrants indexed as Locations 2 and 4). The last two likelihood arrays $\mathbf{A}^{(2),2}$ and $\mathbf{A}^{(2),3}$ map to respective observation modalities $\mathbf{o}^{(2),2}$ and $\mathbf{o}^{(2),3}$, and are not shown for clarity; the $\mathbf{A}^{(2),2}$ likelihood encodes the joint probability of particular types of trial feedback (Null, Correct, Incorrect—encoded by $\mathbf{o}^{(2),2}$) as a function of the current hidden scene and the location of the agent's eyes, while $\mathbf{A}^{(2),3}$ is an unambiguous proprioceptive mapping that signals to the agent the location of its own eyes via $\mathbf{o}^{(2),3}$. Note that these two last observation modalities $\mathbf{o}^{(2),2}$ and $\mathbf{o}^{(2),3}$ are directly sampled from the environment, and are not passed up as "inferred observations" from Level 1.

well as their "spatial permutability" (with respect to quadrant-occupancy) by means of a single hidden state factor that exhaustively encodes the unique combinations of scenes and their spatial configurations. This first hidden state factor is thus a 48-dimensional state distribution (4 scenes × 12 possible spatial configurations—see **Figure 7** for visual illustration).

The second hidden state factor corresponds to the current spatial position that's being visually fixated—this can be thought of as a hidden state encoding the current configuration of the agent's eyes. This hidden state factor has nine possible states: the first state corresponds to an initial position for the eyes (i.e., a fixation region in the center of the array); the next four states (indices 2–5) correspond to the fixation positions of the four quadrants in the array, and the final four states (6–9) correspond to categorization choices (i.e., a saccade which reports the agent's guess about the scene identity). The states of the first and second hidden state factors jointly determine which observation is sampled at each timestep on Level 2.

Observations at this level comprise three modalities. The first modality encodes the identity of the visual stimulus at the fixated location and is identical in form to the first hidden state factor at Level 1: namely, it can be either the "Null" outcome (when there is no visual stimulus at the fixated location) or one of the four motion directions. The likelihood matrix for the first-modality on Level 2, namely $\mathbf{A}^{(2),1}$, consists of probabilistic mappings from the scene identity /spatial configuration (encoded by the first hidden state factor) and the current fixation location (the second hidden state factor) to the stimulus identity at the fixated location, e.g., if the scene is **UP-RIGHT** under the configuration where the **UP**wards-moving RDM is in the upper left quadrant and the **RIGHT**wards-moving RDM is in the upper right quadrant and the current fixation location (the second

hidden state) is the upper left quadrant, then the likelihood function will determine the first-modality observation at Level 2 to be **UP**. When the agent is fixating either an empty quadrant, the starting fixation location, or one of the response options (locations 6–9), the observation in the first modality is **Null**. The likelihood functions are deterministic and identical in both the generative model and generative process—this imbues the agent with a kind of "prior knowledge" of the (deterministic) mapping between the scenes and their respective visual manifestations in the 2×2 grid. The second observation modality is a ternary variable that returns feedback to the agent based on its scene categorization performance—it can assume the values of "No Feedback," "Correct," or "Incorrect." Including this observation modality (and prior beliefs about the relative probability of its different values) allows us to endow agents with the drive to report their guess about the scene, and to do so accurately in order to maximize the chance of receiving correct feedback. The likelihood mapping for this modality $\mathbf{A}^{(2),2}$ is structured to return a "No Feedback" outcome in this modality when the agent fixates any area besides the response options, and returns "Correct" or "Incorrect" once the agent makes a saccade to one of the response options (locations 6–9)—the particular value it takes depends jointly on the true hidden scene and the scene identity that the agent has guessed. We will further discuss how a drive to respond accurately emerges when we describe the prior beliefs parameterized by the **C** and **D** arrays. The final observation modality at Level 2 is a proprioceptive mapping (similar to "sampling-state" modality at Level 1) that unambiguously signals which location the agent is currently fixating via a 9×9 identity matrix $\mathbf{A}^{(2),3}$.

The transition matrices at Level 2, namely $\mathbf{B}^{(2),1}$ and $\mathbf{B}^{(2),2}$, describe the dynamics of the scene identity and of the agent's oculomotor system, respectively. We assume the dynamics that describe the scene identity are both uncontrolled and unchanging, and thus fix $\mathbf{B}^{(2),1}$ to be an identity matrix that ensures the scene identity/spatial configuration is stable over time. As in earlier formulations (Friston et al., 2012a; Mirza et al., 2016, 2019a) we model saccadic eye movements as transitions between control states in the 2nd hidden state factor. The dynamics describing the eye movement from the current location to a new location is encoded by the transition array $\mathbf{B}^{(2),2}$ (e.g., if the action taken is 3 then the saccade destination is described by a transition matrix that contains a row of 1s on the third row, mapping from any previous location to location 3).

Inference and action selection at Level 2 proceeds as follows: based on the current hidden state distribution and Level 1's likelihood mapping $\mathbf{A}^{(1),1}$ (the generative process), observations are sampled from the three modalities. The observation under the first-modality at this level (either "Null" or a motion direction parameterizing an RDM stimulus) is passed down to Level 1 as the initial *true* hidden state. The agent also generates expectations about the first-modality observations via $\mathbf{A}^{(1),1} \cdot Q(\mathbf{s}_t)$, where $\mathbf{A}^{(1),1}$ is the generative model's likelihood and $Q(\mathbf{s}_t)$ is the latest posterior density over hidden states (factorized into scene identity and fixation location). This predictive density over (first-modality) outcomes serves as an *empirical prior* for the agent's beliefs about the hidden states in the first factor—motion direction—at Level 1. Belief-updating and policy selection at Level 1 then proceeds via active inference using the empirical priors inherited from Level 2 in addition to its own generative model and process (as described in Section 4.2.1). Once the motion observations and belief updating terminates at Level 1, the final posterior beliefs about the 1st factor hidden states are passed to Level 2 as "inferred" observations of the first modality. The belief updating at Level 2 proceeds as usual, where observations (both those "inferred" from Level 1 and the true observations from the Level 2 generative process: the oculomotor state and reward modality) are integrated using Level 2's generative model to form posterior beliefs about hidden states and policies. The policies at this level, like at the lower level, only consider one step ahead in the future—so each policy consists of one action (a saccade to one of the quadrants or a categorization action), to be taken at the next timestep. An action is sampled from the posterior over policies $a_t \sim Q(\pi)$, which changes hidden states in the next time step to generate a new observation, thus closing the action-perception cycle. In this spatiotemporally "deep" version of scene construction, we see how a temporally-extended process of active inference at the lower level (capped at $\mathbf{T}_1 = 20$ time steps in our case) can be nested within a single time step of a higher-level process, endowing such generative models with a flexible, modular form of temporal depth. Also note the asymmetry in informational scheduling across layers, with posterior beliefs about those hidden states linked with the higher level being passed *up* as evidence for outcomes at the higher level, with observations at the higher level being passed *down* as empirical priors over hidden states at the lower level.

4.2.3. Priors

In addition to the likelihood **A** and **B** arrays that prescribe the probabilistic relationships between variables at each level, the generative model is also equipped with prior beliefs over observations and hidden states that are respectively encoded in the so-called **C** and **D** arrays. See **Figure 8** for schematic analogies for these arrays and their elements for the two hierarchical levels.

The **C** array contains what are often called the agent's "preferences" $P(o)$ and encodes the agent's prior beliefs about observations (an unconditional probability distribution). Rather than an explicit component of the generative model, the prior over outcomes is absorbed into the prior over policies $P(\pi)$, which is described in Section 2.2. Policies that are more likely to yield observations that are deemed probable under the prior (expressed in terms of agent's preferences $P(o)$) will have less expected free energy and thus be more likely to be chosen. *Instrumental value* or expected utility measures the degree to which the observations expected under a policy correlate with prior beliefs about those observations. For categorical distributions, evaluating instrumental value amounts to taking the dot product of the (policy-conditioned) posterior predictive density over observations $Q(o_\tau|\pi)$ with the log probability density over outcomes $\log P(o_\tau)$. This reinterpretation of preferences as prior beliefs about observations allows us to discard the classical notion of a "utility function" as postulated in fields like reward neuroscience and economics, instead explaining both epistemic and instrumental behavior using the

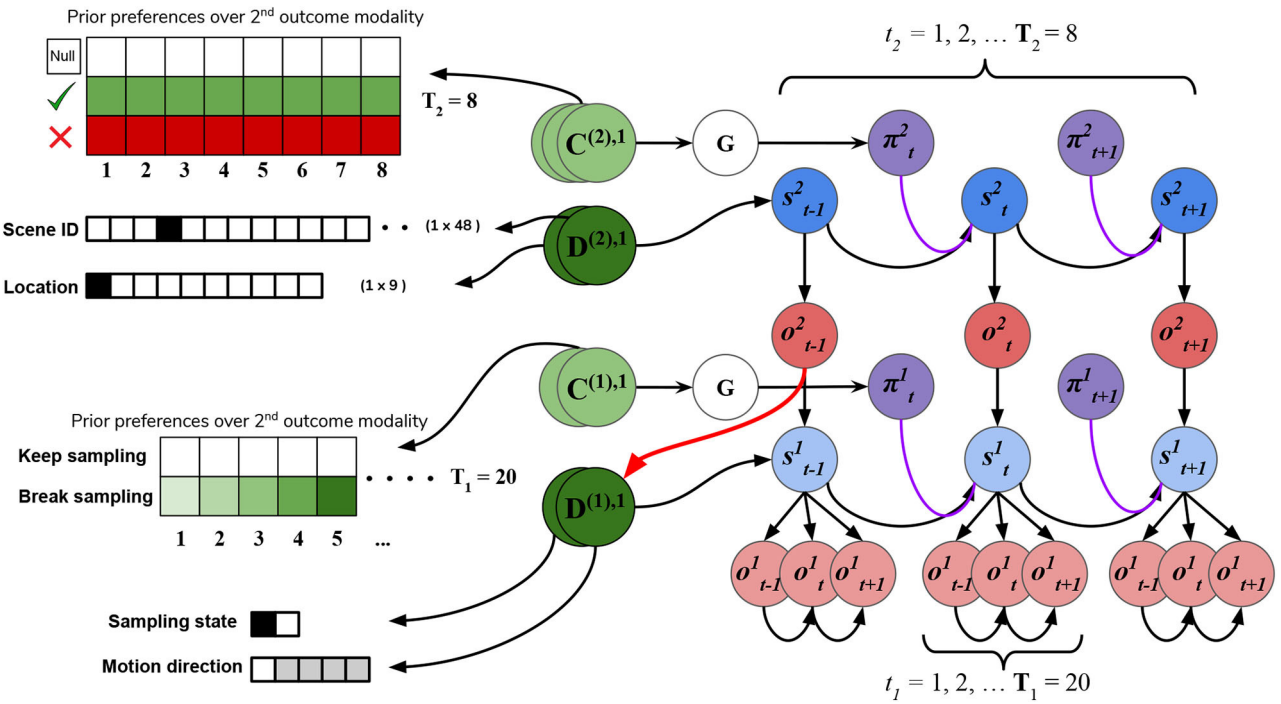


FIGURE 8 | C's and **D**'s. Prior beliefs over observations and hidden states for both hierarchical levels. Note that superscripts here index the hierarchical level, and separate modalities/factors for the **C** and **D** matrices are indicated by stacked circles. At the highest level (Level 2), prior beliefs about second-modality outcomes ($\mathbf{C}^{(2),2}$) encode the agent's beliefs about receiving correct and avoiding incorrect feedback. Prior beliefs over the other outcome modalities ($\mathbf{C}^{(2),1}$ and $\mathbf{C}^{(2),3}$) are all trivially zero. These beliefs are stationary over time and affect saccade selection at Level 2 via the expected free energy of policies **G**. Prior beliefs about hidden states $\mathbf{D}^{(2)}$ at this level encode the agent's initial beliefs about the scene identity and the location of their eyes. This prior over hidden states can be manipulated to put the agent's beliefs about the world at odds with the actual hidden state of the world. At Level 1, the agent's preferences about being in the "**Break-sampling**" state increases over time and is encoded in the preferences about second modality outcomes ($\mathbf{C}^{(1),2}$), which corresponds to the agents umambiguous perception of its own sampling state. Finally, the prior beliefs about initial states at Level 1 ($\mathbf{D}^1$) correspond to the motion direction hidden state (the RDM identity) as well as which sampling-state the agent is in. Crucially, the first factor of these prior beliefs $\mathbf{D}^{(1),1}$ is initialized as the "expected observations" from Level 2: the expected motion direction (first modality). These expected observations are generated by passing the variational beliefs about the scene at Level 2 through the modality-specific likelihood mapping: $Q(o^{(2),1}|s^{(2),1}) = P(o^{(2),1}|s^{(2),1})Q(s^{(2),1})$. The prior over hidden states at Level 1 is thus called an *empirical prior* as it is inherited from Level 2. The red arrow indicates the relationship between the expected observation from Level 2 and the empirical prior over (first-factor) hidden states at Level 1.

common currency of log-probabilities and surprise. In order to motivate agents to categorize the scene, we embed a self-expectation of accuracy into the **C** array of Level 2; this manifests as a high prior expectation of receiving "Correct" feedback (a relative log probability of +2 nats) and an expectation that receiving "Incorrect" feedback is unlikely (relative log probability of −4 nats). The remaining outcomes of the other modalities at Level 2 have equal log-probability in the agent's prior preferences, thus contributing identically (and uninformatively) to instrumental value. At Level 1 we encoded a form of "urgency" using the **C** matrix; we encoded the prior belief that the probability of observing the "**Break-sampling**" state (via the umambiguous mapping $A^{(1),2}$) increases over time. This necessitates that the complementary probability of remaining in the "**Keep-sampling**" state decreases over time. Equipping the Level 1 MDP with such preferences generates a tension between the epistemic drive to resolve uncertainty about the hidden state of the currently-fixated stimulus and the ever-strengthening prior preference to terminate sampling at Level 1. In the simulation results to follow, we explore this tension more explicitly and report an interesting yet unexpected relationship between sensory uncertainty and fixational dwell time, based on the dynamics of various contributions to expected free energy.

Finally, the **D** array encodes the agent's initial (prior) beliefs over hidden states in the environment. By changing prior beliefs about the initial states, we can manipulate an agent's beliefs about the environment independently of the true hidden states characterizing that environment. In the Section 5.2 below we describe the way we parameterize the first hidden state factor of the Level 2 **D** matrix to manipulate prior beliefs about the scene. The second hidden state factor at Level 2 (encoding the saccade location) is always initialized to start at Location 1 (the generic "starting" location). At Level 1, the first-factor of the **D** matrix (encoding the true motion direction of an RDM) is initialized to the posterior expectations from Level 2, i.e., $Q(o^{(1),1}|\mathbf{s}_t) = \mathbf{A}^{(1),1}Q(\mathbf{s}_t)$. The second-factor belief about hidden states (encoding the sampling state) is initialized to the "**Keep-sampling**" state.

In the following sections, we present hierarchical active inference simulations of scene construction, in which

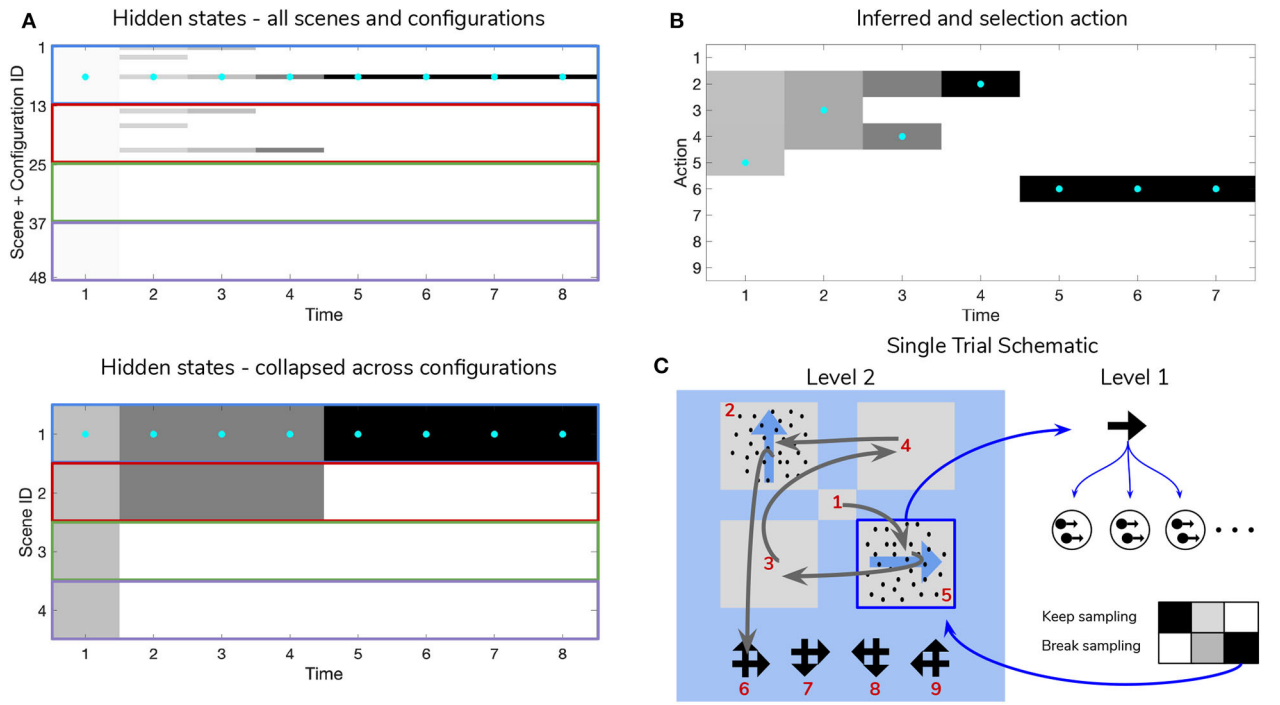


FIGURE 9 | Simulated trial of scene construction under high sensory precision. **(A)** The evolution of posterior beliefs about scene identity—the first factor of hidden states at Level 2—as a deep active inference agent explores the visual array. In this case, sensory precision at Level 1 is high, meaning that posterior beliefs about the motion direction of each RDM-containing quadrant are resolved easily, resulting in fast and accurate scene categorization. Cells are gray-scale colored according to the probability of the belief for that hidden state and time index (darker colors correspond to higher probabilities). Cyan dots indicates the true hidden state at each time step. The top row of **(A)** shows evolving beliefs about the fully-enumerated scene identity (48 possibilities), with every 12 configurations highlighted with a differently-colored bounding box, correspond to beliefs about each type of scene (i.e., **UP-RIGHT**, **RIGHT-DOWN**, **DOWN-LEFT**, **LEFT-UP**). The bottom panel shows the collapsed beliefs over the four scenes, computed by summing the hidden state beliefs across the 12 spatial configurations. **(B)** Evolution of posterior beliefs about actions (fixation starting location not shown), culminating in the categorization decision (here, the scene was categorized as **UP-RIGHT**, corresponding to a saccade to location 6. **(C)** Visual representation of the agent's behavior for this trial. Saccades are depicted as curved gray lines connecting one saccade endpoint to the next. Fixation locations (corresponding to 2nd factor hidden state indices) are shown as red numbers. The Level 1 active inference process occurring within a single fixation is schematically represented on the right side, with individual motion samples shown as issued from the true motion direction via the low level likelihood $\mathbf{A}^{(1),1}$. The agent observes the true RDM at Level 1 and updates its posterior beliefs about this hidden state. As uncertainty about the RDM direction is resolved, the "**Break-sampling**" action becomes more attractive (since epistemic value contributes increasingly less to the expected free energy of policies). In this case, the sampling process at Level 1 is terminated after only three timesteps, since the precision of the likelihood mapping is high ($p = 5.0$) which relates to the speed at which uncertainty is resolved about the RDM motion direction—see the text for more details.

we manipulate the uncertainty associated with beliefs at different levels of the generative model to see how uncertainty differentially affects inference across levels in uncertain environments.

5. SIMULATIONS

Having introduced the hierarchical generative model for our RDM-based scene construction task, we will now explore behavior and belief-formation in the context of hierarchical active inference. In the following sections we study different aspects of the generative model through quantitative simulations. We relate parameters of the generative model to both "behavioral" read-outs (such as sampling time, categorization latency and accuracy) as well as the agents' internal dynamics (such as the evolution of posterior beliefs, the contribution of different kinds of value to policies, etc.). We then discuss the implications of our model for studies of hierarchical inference in noisy, compositionally-structured environments.

5.1. Manipulating Sensory Precision

Figures 9, **10** show examples of deep active inference agents performing the scene construction task under two levels of motion coherence (high and low, respectively for **Figures 9**, **10**), which is equivalent to the reliability of motion observations at Level 1. In particular, we operationalize this uncertainty via an inverse temperature p that parameterizes a softmax transformation on the columns of the Level 1 likelihood mapping to RDM observations $\mathbf{A}^{(1),1}$. Each each column of $\mathbf{A}^{(1),1}$ is initialized as a "one-hot" vector that contains a probability of 1 at the motion observation index corresponding to the true motion direction, and 0s elsewhere. As p decreases, $\mathbf{A}$ deviates further from the identity matrix and Level 1 motion observations become more degenerate with respect to the hidden state (motion direction) underlying them. Note that this parameterization of

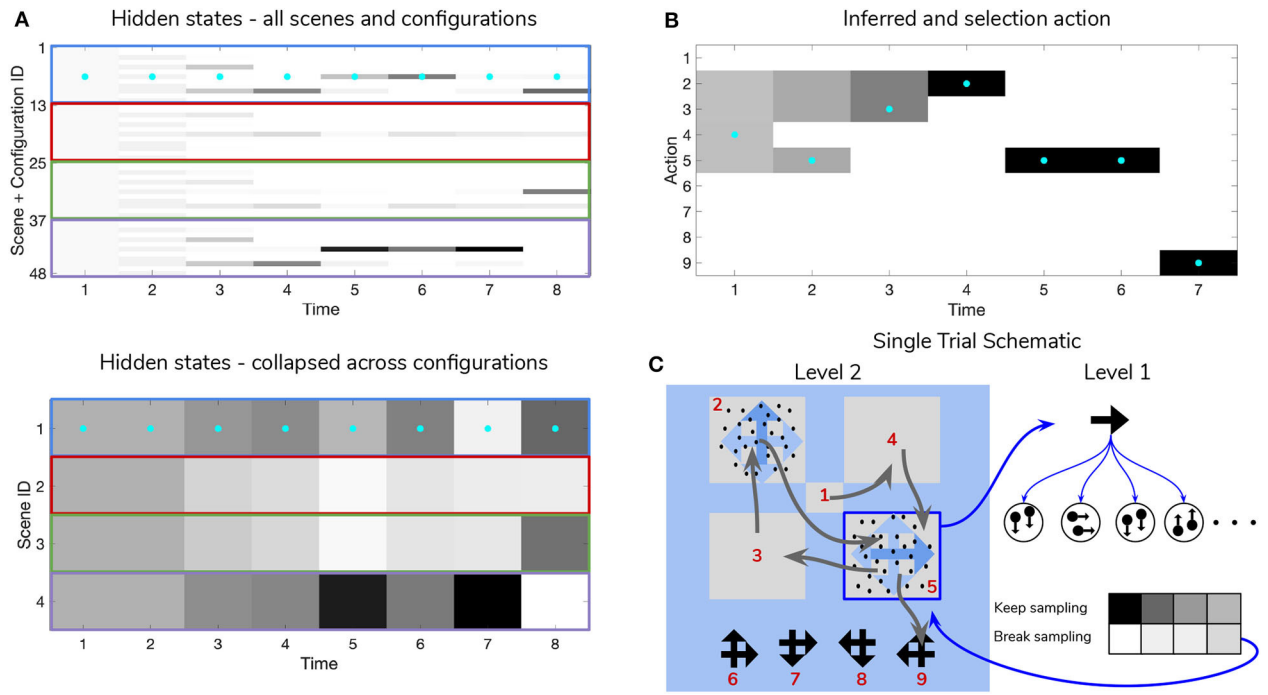


FIGURE 10 | Simulated trial of scene construction with low sensory precision. Same as in **Figure 9**, except in this trial the precision of the mapping between RDM motion directions and samples thereof is lower, $p = 0.5$. This leads to an incorrect sequence of inferences, where the agent ends up believing that the scene identity is **LEFT-UP** and guessing incorrectly. Note that after this choice is made and incorrect feedback is given, the agent updates their posterior in terms of the "next best" guess, which is from the agent's perspective either **UP-RIGHT** or **DOWN-LEFT** (see the posterior at Time step 8 of **(A)**). **(C)** Shows that the relative imprecision of the Level 1 likelihood results in a sequence of stochastic motion observations that frequently diverge from the true motion direction (in this case, the true motion direction is **RIGHT** in the lower right quadrant (Location 5)). Level 1 belief-updating gives rise to an imprecise posterior belief over motion directions that are passed up as inferred outcomes to Level 2, leading to false beliefs about the scene identity. Note the "ambivalent," quadrant-revisiting behavior, wherein the agent repeatedly visits the lower-right quadrant to resolve uncertainty about the RDM stimulus at that quadrant.

motion incoherence only pertains to the last four rows/columns of $\mathbf{A}^{(1),1}$, as the first row/column of the likelihood ($\mathbf{A}^{(1),1}(1,1)$) corresponds to observations about the "Null" hidden state, which is always observed unambiguously when it is present. In other words, locations that do not contain RDM stimuli are always perceived as "Null" in the first modality with certainty.

Figure 9 is a simulated trial of scene construction with sensory uncertainty at the lower level set to $p = 5.0$. This manifests as a stream of motion observations at the lower level that reflect the true motion state $\sim$ 98% of the time, i.e., highly-coherent motion. As the agent visually interrogates the 2×2 visual array (the 2^{nd} to 5^{th} rows of Panel **B**), posterior beliefs about the hidden scene identity (Panel **A**) converge on the true hidden scene. After the first RDM in the lower right quadrant is seen (and its state resolved with high certainty), the agent's Level 2 posterior starts to only assign non-zero probability to scenes that include the **RIGHT**wards-moving motion stimulus. Once the second, **UP**wards-moving RDM stimulus is perceived in the upper left, the posterior converges upon the correct scene (in this case, indexed as state 7, one of the 12 configurations of **UP-RIGHT**). Once uncertainty about the hidden scene is resolved, **G** becomes dominated by instrumental value, or the dot-product of counterfactual observations with prior preferences. Expecting to receive correct feedback, the agent saccades to location 6 (which corresponds to the scene identity **UP-RIGHT**) and receives a "Correct" outcome in the second-modality of Level 2 observations. The agent thus categorizes the scene and remains there for the remainder of the trial to exploit the expected instrumental value of receiving "Correct" feedback (for the discussion about how behavior changes with respect to prior belief and sensory precision manipulations, we only consider behavior up until the time step of the first categorization decision).

Figure 10 shows a trial when the RDMs are incoherent ($p = 0.5$, meaning the Level 1 likelihood yields motion observations that reflect the true motion state $\sim$35% of the time). In this case, the agent fails to categorize the scene correctly due to the inability to form accurate beliefs about the identity of RDMs at Level 1—this uncertainty carries forward to lead posterior beliefs at Level 2 astray. Interestingly, the agent still forms relatively confident posterior beliefs about the scene (see the posterior at Timestep 7 of **Figure 11A**), but they are inaccurate since they are based on inaccurate posterior beliefs inherited from Level 1. This is because even though the low-level belief is built from noisy observations, posterior probability ends up still "focusing" on a particular dot direction based on the particular

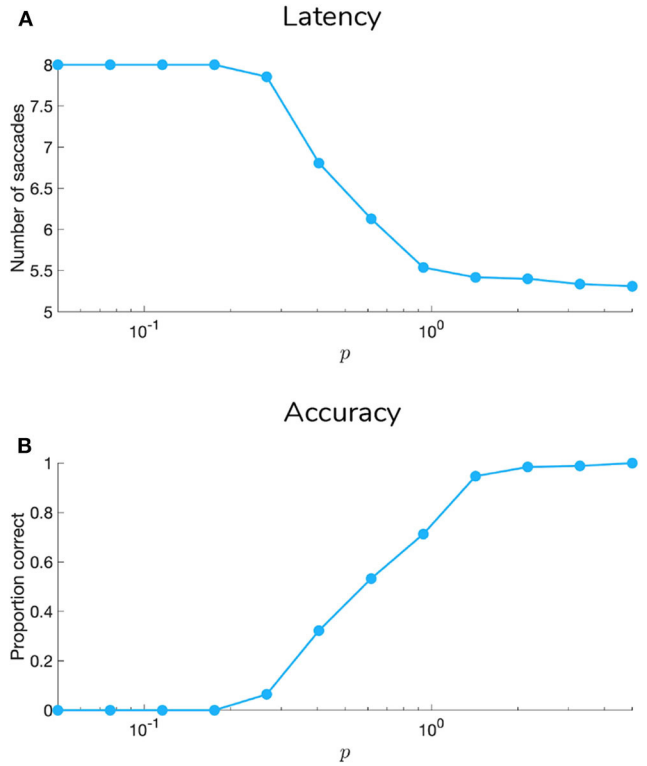


FIGURE 11 | Effect of sensory precision on scene construction performance. Average categorization latency **(A)** and accuracy **(B)** as a function of sensory precision p which controls the entropy of the (Level 1) likelihood mapping from motion direction to motion observation. We simulated 185 trials of scene construction under hierarchical active inference for each level of p (12 levels total), with scene identities and configurations randomly initialized for each trial. Sensory precision is shown on a logarithmic scale.

sequence of observations that is sampled; this is then integrated with empirical priors and subsequent observations to narrow the possible space of beliefs about the scene. The posterior uncertainty also manifests as the time spent foraging in quadrants before making categorization (nearly double the time spent by the agent in **Figure 9**). The cause of this increase in foraging time is 2-fold. First of all, since uncertainty about the scene identity is high, the epistemic value of policies that entail fixations to RDM-containing quadrants remains elevated, even after all the quadrants have been visited. This is because uncertainty about hidden states is unlikely to be resolved after a single saccade to a quadrant with an incoherent RDM, meaning that the epistemic value of repeated visits to such quadrants decreases slowly with repeated foraging. Secondly, since Level 2 posterior beliefs about the scene identity are uncertain and are distributed among different states, the instrumental value of categorization actions remains low—remember that instrumental value depends not only on the instrumental value of receiving "Correct" feedback, but also on the agent's expectation about the *probability* of receiving this feedback upon making an action, relative to the probability of receiving "Incorrect" feedback. The relative values of the prior preferences for being "Correct" vs. "Incorrect" thus tune the risk-averseness of the agent, and manifest as a dynamic balance between epistemic and instrumental value. See Mirza et al. (2019b) for a quantitative exploration of these prior preferences and their effect on active inference.

We quantified the relationship between sensory precision and scene construction performance by simulating scene construction trials under different sensory precisions p (see **Figure 11**). The two measures shown are: (1) *categorization latency* (**Figure 11A**), defined as the number of time steps elapsed before a saccade to one of the choice locations is initiated; and (2) *categorization accuracy* (**Figure 11B**), defined as percentage of trials when the agent's first categorization resulted in "Correct" feedback. In agreement with intuition, for low values of p agents take more time to categorize the scene and categorize less accurately. As sensory precision increases, agents require monotonically less time to forage the array before categorizing, and this categorization also becomes more accurate. In the next section, we explore the relationship between sensory precision and performance when the agent entertains prior beliefs of varying strength about the probability of a certain scene.

5.2. Manipulating Prior Beliefs

For the simulations discussed in the previous section, agents always start scene construction trials with "flat" prior beliefs about the scene identity. This means that the first factor of the prior beliefs about hidden states at Level 2 $\mathbf{D}^{(2),1}$ was initialized as a uniform distribution. We can manipulate the agent's initial expectations about the scenes and their spatial arrangements by arbitrarily sculpting $\mathbf{D}^{(2),1}$ to have high or low probabilities over any state or set of states. Although many manipulations of the Level 2 prior over hidden states are possible, here we introduce a simple prior belief manipulation by uniformly elevating the prior probability of all spatial configurations (12 total) of a single type of scene. For example, to furnish an agent with the belief that there's a 50% chance of any given trial being a **RIGHT-DOWN** scene, we simply boost the probabilities associated with hidden states 13–25 (the 12 spatial configurations of the **RIGHT-DOWN** scene) relative to the other hidden scenes, so that the total integrated probability of hidden states 13–25 is 0.5. This implies that the other hidden scenes each now have $\frac{(1-0.5)}{3} \approx 0.1667$ probability, once respectively integrated over their 12 configuration states. **Figure 12** shows the effect of parametrically varying the strengths of prior beliefs on the same behavioral measures shown in **Figure 11**. Similar to **Figures 11**, **12** demonstrates a monotonic increase in accuracy with increasing sensory precision, regardless of how much the agent initially expects a particular scene type. This means that strong but incorrect prior beliefs (over initial states) can still be "overcome" with reliable enough sensory data. However, agents with stronger priors are less sensitive to the increase in sensory precision than their "flat-priored" counterparts, as can be seen by the lower accuracy level of the most purple-colored lines in **Figure 12**. Note that the averages shown are only for agents with "incorrect" prior beliefs; namely, the prior over hidden states in the generative model for each trial was always initialized to be a different scene type than the true scene. This has the effect of setting the minimum accuracy for the "strongest-priored"

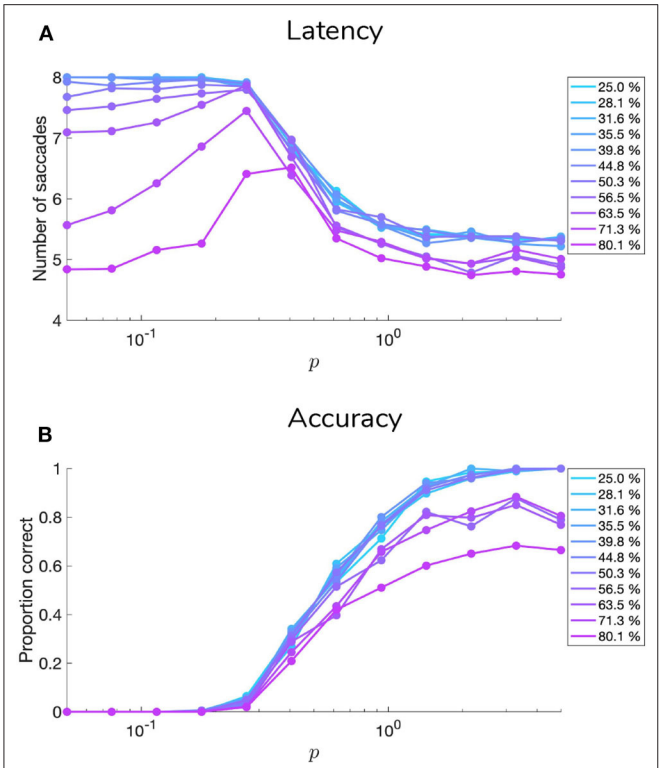


FIGURE 12 | Effect of sensory precision on scene construction performance for different prior belief strengths. Same as in **Figure 11** but for different strengths of initial prior beliefs (legend on right). Prior belief strengths are defined as the probability density of the prior beliefs about hidden states (1st hidden state factor of Level 2—$\mathbf{D}^{(2),1}$) concentrated upon one of the four possible scenes. This elevated probability is uniformly spread among the 12 hidden states corresponding to the different quadrant-configurations of that scene, such that the agent has no prior expectation about a particular arrangement of the scene, but rather about that scene type in general. Here, we only show the results for agents with "incorrect" prior beliefs—namely, when the scene that the agent believes to be at play is different from the scene actually characterizing the trial.

agents (who typically categorize the scene identity at the first time step) at 0% rather than 25% (chance performance). These results are consistent with the fundamental relationship between the likelihood term and prior probability in Bayes' theorem (see Equation 1): the posterior over hidden states is calculated as the product of the likelihood and the prior. Increasing the precision of one of these two will "shift" the posterior distribution in the respective direction of the more precise distribution. This manifests as a parametric "de-sensitizing" of posterior beliefs to sensory evidence as priors become stronger. This balance between sensory and prior precision is exactly manifested in the prior-dependent sensitivity of the accuracy curves in **Figure 12B**.

The interaction between sensory and prior precision is not as straightforward when it comes to categorization latency. **Figure 12A** shows that when the sensory precision p is high enough, most of the variance in latency introduced by prior beliefs vanishes, since observations alone can be relied on to ensure fast inference about the scene. For low values of p, however, latency is highly-sensitive to prior belief strength. Under weak prior beliefs and low p, the agent displays ambivalence—beliefs about RDM direction at Level 1 are not precise enough to enable scene inference, causing the agent to choose the policies that have (albeit) small epistemic value while avoiding the risk of categorizing incorrectly. This causes the agent to saccade among RDM-containing quadrants. Agents with stronger prior beliefs, however, do not rely on observations to determine posterior beliefs because their prior beliefs about the scene already lend high instrumental value to categorization actions. This corresponds to trials when the agent categorizes the scene immediately (for the strongest prior beliefs, this occurs even before inspecting any quadrants) and relying minimally on sensory evidence. This faster latency comes at the cost of accuracy, however, as evident from the lower average accuracy of strongly-priored agents displayed in **Figure 12B**.

Now we explore the effects of sensory and prior precision on belief-updating and policy selection at the lower level, during a single quadrant fixation. **Figure 13A** shows the effect of increasing p on the break-time (or to analogize it more directly to eye movements: the fixational "dwell time") at Level 1. We observe a non-trivial, inverted-U relationship between the logarithm of p (our analog of motion coherence) and the time it takes for agents to break the sampling at Level 1. For the lowest (most incoherent) values of the likelihood precision p, the agents dwell for as little time as they do as for the highest precisions. Understanding this paradoxical effect requires a more nuanced understanding of epistemic value. In general, increasing the precision of the likelihood mapping increases the amount of uncertainty that observations can resolve about hidden states, thus lending high epistemic value to policies that disclose such observations (Parr and Friston, 2017). An elevated epistemic value predicts an increase in dwell time (i.e., via an increase in the epistemic value for the "**Keep-sampling**" policy at Level 1) for increasing sensory precision. However, an increased precision of the Level 1 likelihood also implies that posterior uncertainty is resolved at a faster rate (due to high mutual information between observations and hidden states), which suppresses epistemic value over time. The rate at which epistemic value drops off thus increases in the presence of informative observations, since the posterior converges to a tight probability distribution relatively quickly. On the other hand, at very low likelihood precisions, the low information content of observations in addition to the linearly-increasing cost of sampling (encoded in the Level 1 preferences $\mathbf{C}^{(1),2}$) renders the sampling of motion observations relatively useless for agents, and it "pays" to just break sampling early. This results in the pattern of break-times that we observe in **Figure 13A**.

It is worth mentioning the barely noticeable effect of prior beliefs (**Figure 13A**) about the scene identity on break times at Level 1. Although prior beliefs about the scene at Level 1 manifest as empirical priors over hidden states (motion directions) at Level 2, it seems that the likelihood matrix plays a much larger role in determining break times than the initial beliefs. This means that even when the agent initially assigns relatively more probability to particular RDM directions (conditional on beliefs about scenes at Level 2), this initial belief can quickly

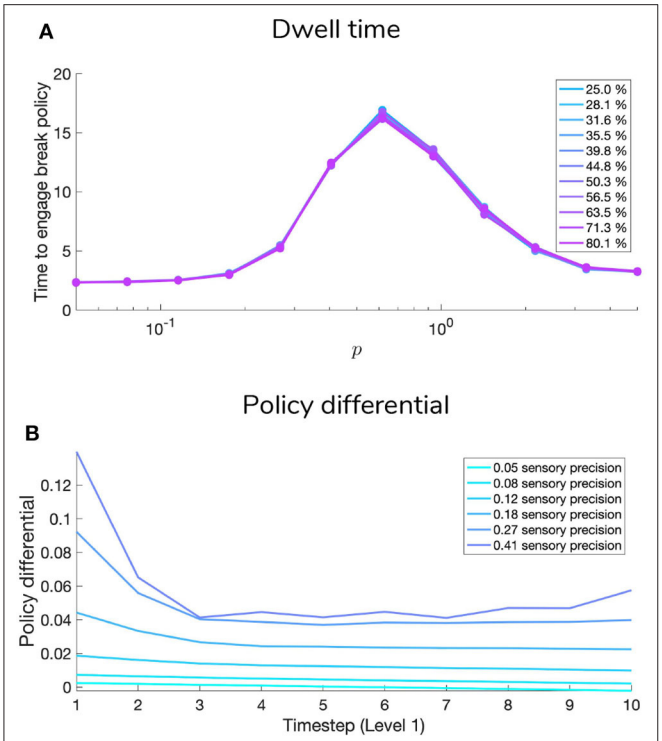


FIGURE 13 | Effect of sensory precision on quadrant dwell time. **(A)** Shows the effect of increasing sensory precision at Level 1 on the time it takes to switch to "**Break-sampling**" policy. Here, 250 trials were simulated for each combination of sensory precision and prior belief strength, with priors over hidden states at Level 2 randomly initialized to have high probability about 1 of the 4 scene types. Break-times were analyzed only for the first saccade (at Level 2) of each trial. **(B)** Shows the effect of sensory precision on evolution of the relative posterior probabilities of the "**Keep-sampling**" vs. the "**Break-sampling**" policies (*Policy Differential* = $P_{\text{Keep-sampling}} - P_{\text{Break-sampling}}$). We only show these posterior policy differentials for the first 10 time steps of sampling at Level 1 due to insufficient numbers of saccades that lasted more than 10 time steps at the highest/lowest sensory precisions (see **A**). Averages are calculated across different prior belief strengths, based on the lack of an effect, as is apparent in **(A)**. The policy differential defined in this way is always positive because as soon as the probability of "**Break-sampling**" exceeds that of "**Keep-sampling**" (i.e., *Policy Differential* < 0), the "**Break-sampling**" policy will be engaged with near certainty. This is due the high precision over policies at the lower level (here, $\gamma = 512$), which essentially ensures that the policy with higher probability will always be selected.

be revised in light of incoming evidence (namely, observations at Level 1, inverted through the likelihood mapping to produce a marginal posterior over hidden states). This also speaks to the segregation of belief-updating between hierarchical levels; although beliefs about hidden states and observations are passed up and down the hierarchy, belief-updating occurs only with respect to the variational free energy of a particular layer's generative model, thus insulating variational updating to operate at distinct spatiotemporal scales. This results in the conditional independence of decision-making across hierarchical levels, and clarifies the dissociable influence of prior about scenes on Level 1 vs. Level 2. For example, even on trials when an agent has strong prior beliefs about the scene and thus takes fewer saccades to categorize it, differences in lower-level "dwell time" are still largely determined by the sensory precision p of the likelihood mapping and the preference to enter the "**Break-sampling**" state, encoded as an increasing probability to observe oneself occupying this state (in $\mathbf{C}^{(1),2}$).

The curves in **Figure 13B** clarify the rate at which epistemic value decreases for high sensory precisions. The "policy differential" measures the difference between the posterior probability of the "**Keep-sampling**" vs. "**Break-sampling**" policies at Level 1: $P_{\text{Keep-sampling}} - P_{\text{Break-sampling}}$. At the lowest sensory precisions, there is barely any epistemic value to pursuing the "**Keep-sampling**" policy, allowing the break policy to increasingly dominate action-selection over time. For higher sensory precisions, the "**Keep-sampling**" policy starts with >10% more probability than the "**Break-sampling**" policy since the epistemic value of sampling observations is high, but quickly loses its advantage as posterior uncertainty is resolved. At this point the probability of breaking becomes more probable, since posterior beliefs about the RDM are fairly resolved and the instrumental of breaking is only getting higher with time.

6. DISCUSSION

In the current work, we presented a hierarchical partially-observed Markov Decision Process model of scene construction, where scenes are defined as arbitrary constellations of random dot motion (RDM) stimuli. Inspired by an earlier model of scene construction (Mirza et al., 2016, 2018) and a deep temporal formulation of active inference (Friston et al., 2017d), we cast this scene construction task as approximate Bayesian inference occurring across two hierarchical levels of inference. One level involves optimizing beliefs about the instantaneous contents of agent-initiated visual fixations; the second level involves integrating the contents of different fixated locations to form beliefs about a higher-level concept like a scene. Through simulations we showed how this deep, temporal model formulation can be used to provide an active inference account of behavior in such compositional inference tasks. Deep active inference agents performing scene construction exhibit the Bayesian hallmarks of a dynamic trade-off between sensory and prior precision when it comes to scene inference and saccade selection. The hierarchical segregation of inference between saccadic and fixational levels gives rise to unexpected effects of sensory uncertainty at the level of single fixations, where we observe an inverted-U relationship between motion coherence and fixational dwell time. This non-linear relation can be explained by appealing to the evolution of epistemic value over time, under the assumption that the agent entertains beliefs about the precision of the environmental process generating visual sensations, while simultaneously optimizing the sufficient statistics of beliefs about the currently-fixated stimulus. The fact that the precision of the likelihood mapping increases the epistemic value of policies that furnish observations sampled from the generative process, while simultaneously increasing the rate at which posterior uncertainty is reduced, explains

the non-monotonic influence of sensory precision on Level 1 decision latency.

These results contrast with the predictions of classic evidence accumulation models like the drift-diffusion model or DDM (Ratcliff, 1978; Palmer et al., 2005; Ratcliff and McKoon, 2008). In the drift-diffusion model, reaction times are modeled as proportional to the latency it takes for a time-varying decision variable (or **DV**) to reach one of two fixed decision boundaries **Z** and −**Z** that respectively correspond to two hypotheses (e.g., the equivalent of sufficiently-strong posterior beliefs in one of two hidden states). At each time step, increments to the **DV** are calculated as the log of the ratio between the evidence for each hypothesis conditioned on observations. In discrete-time environments this update-rule for **DV** is equivalent to the Sequential Probability Ratio Test formulated by Wald and Wolfowitz (1948). For time-independent decision boundaries and a fixed initial value of the **DV**, a drift-diffusion process yields a monotonic decreasing relationship between motion incoherence and decision latency (Bogacz et al., 2006; Ratcliff and McKoon, 2008), where motion coherence factors into the DDM as the drift rate of the **DV**—this is analogous to the *sensitivity* of the **DV** to incoming sensory evidence. In the current active inference model, we have binarized policies at Level 1 in part to invite comparison between our model and DDM models (which in their classical form handle binary hypotheses). Rather than modeling actions as discrete perceptual decisions about the most likely hidden state underlying observations (since in the current context, we have a 4-dimensional RDM state space), we instead model the decision as selecting between one of two "sampling" policies, whose probabilities change over time due to the dynamics of the expected free energy. This evolving action-probability weighs epistemic drives to resolve uncertainty against prior preferences that encode an increasing "urgency" to break sampling. This parameterization of decision-making permits a flexible (and in this case, somewhat unexpected) relationship between sensory uncertainty and decision latency (see **Figure 13**). We thus provide a novel, principled prediction for the relationship between sensory uncertainty and reaction time at different levels of inference in perceptual decision-making tasks.

A discussion of the relationship between the current model and previous hierarchical POMDP schemes is also warranted. The model most closely related to the current work is the "deep temporal model" of active reading, proposed by Friston et al. (2017d); the inference schemes are identical, with the critical difference being the way in which updating is terminated at the lower level. In Friston et al. (2017d), policies at the lower level are driven purely by epistemic value and terminate as a result of posterior uncertainty being reduced beyond a certain pre-determined level. In contrast, the current model introduces an additional "Break-policy" (and corresponding observations of a "Sampling-state") at the lower level, whose selection is used to terminate the Level 1 POMDP. This also allows us to motivate decision-making at the lower level MDP using individual costs or goals, as encoded via the "sampling cost" in the lower level prior over observations $P(o)$, explicitly pitting the epistemic drive to resolve uncertainty about the currently-fixated RDM stimulus against the increasing cost of continuing to fixate. Qualitatively, we found that this leads to a smoother relationship between sensory uncertainty (inverse precision of the Level 1 **A** matrix) and the latencies to engage the break policy ("reaction times"), allowing easier comparison of the current model to other evidence accumulation schemes (e.g., drift-diffusion models).

Insight from the robotics and probabilistic planning literature could also be integrated with the current work to extend deep active inference in its scope and flexibility. For instance, the framework of "planning to see" proposed in Sridharan et al. (2010) can be used to drive selective visual processing of goal-relevant features in the sensorium, an important context-sensitive aspect of visual processing (selective and feature-based attention) that is lacking in the current formulation. Mirza et al. (2019a) introduces an active inference model of selective attention in a visual foraging task; the approach proposed therein might be combined with a hierarchical scheme to generate a fully hierarchical model with goal-driven attention operating at multiple levels.

The hierarchical active inference scheme could also be extended to dynamic environments, where the scene itself changes, either due to intrinsic stochasticity or as a function of the agent's (or other agents') actions. This could simply be changed by encoding appropriate self-initiated state-changes into the transition model (the "B" matrices) or by introducing intrinsic, non-agent-controlled dynamics into the generative process. Ongoing work in the robotics and planning literature has highlighted the challenges of dynamic, structured environments and proposed various schemes to both plan actions and form probabilistic beliefs in such tasks (Ognibene and Demiris, 2013; Ognibene and Baldassare, 2014). Future research might find ways to meaningfully integrate existing approaches from the hierarchical planning and POMDP literature with deep active inference models, such as the one proposed here.

In future investigations, we plan to estimate the parameters of hierarchical active inference models from experimental data of human participants performing a scene construction task, where the identities of visual stimuli are uncertain (the equivalent of manipulating the sensory likelihood at Level 1 of the hierarchy). Data-driven inversion of a deep scene construction model can then be used to explain inter-subject variability in aspects of hierarchical inference behavior as different parameterizations of subject-specific generative models.

DATA AVAILABILITY STATEMENT

The data used in this study are the results of numerical simulations, and as such, we do not provide datasets. The software used to simulate the data and generate associated figures are based on visual foraging and scene construction demos included in SPM v12.0, and can be freely downloaded from https://www.fil.ion.ucl.ac.uk/spm/ as part of the DEM toolbox. The particular versions of these scripts used to implement the deep hierarchical version are available upon request from the authors.

AUTHOR CONTRIBUTIONS

RH and AP conceived the original idea for the project. RH and MM conceived the hierarchical active inference model. RH, AP, and IK designed the scene construction task using random dot motion. MM, TP, and KF gave the critical insight into formulation of the model. RH conducted the simulations and analyzed the results. All authors contributed to the writing of the manuscript.

FUNDING

This work was supported by an ERC Starting Grant (no: 716846) to AP and a seed fund grant from Leibniz ScienceCampus Primate Cognition, Göttingen, Germany to IK and AP. MM (a Perception and Action in Complex Environments member) was supported by the European Union's Horizon 2020 (Marie Sklodowska-Curie Grant 642961). TP was supported by the Rosetrees Trust (Award Number 173346). KF was funded by a Wellcome Trust Principal Research Fellowship (Ref: 088130/Z/09/Z).

ACKNOWLEDGMENTS

The authors would like to thank Brennan Klein for extensive feedback on the paper and the figures, and Brennan Klein and Alec Tschantz for discussion and initial conceptualization of the project. The authors also thank Kai Ueltzhöffer for feedback on the project and discussions relating active inference to drift-diffusion models. The authors also thank the Monash University Network of Excellence for supporting the workshop Causation and Complexity in the Conscious Brain (Aegina, Greece 2018) at which many of the ideas related to this project were developed.

REFERENCES

Bastos, A., Usrey, W., Adams, R., Mangun, G., Fries, P., and Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Beal, M. J. (2004). *Variational algorithms for approximate bayesian inference* (Ph.D. thesis), Gatsby Unit, University College London, London, United Kingdom.

Biehl, M., Guckelsberger, C., Salge, C., Smith, S. C., and Polani, D. (2018). Expanding the active inference landscape: more intrinsic motivations in the perception-action loop. *Front. Neurorobot.* 12:45. doi: 10.3389/fnbot.2018.00045

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112, 859–877. doi: 10.1080/01621459.2017.1285773

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *J. Math. Psychol.* 76, 198–211. doi: 10.1016/j.jmp.2015.11.003

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* 113, 700–765. doi: 10.1037/0033-295X.113.4.700

Felleman, D. J., and Van, D. E. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Ferro, M., Ognibene, D., Pezzulo, G., and Pirrelli, V. (2010). Reading as active sensing: a computational model of gaze planning during word recognition. *Front. Neurorobot.* 4:6. doi: 10.3389/fnbot.2010.00006

Feynman, R. (1998). *Statistical Mechanics: A Set of Lectures (Advanced Book Classics)*. Boulder, CO: Westview Press.

FitzGerald, T. H. B., Dolan, R. J., and Friston, K. (2015). Dopamine, reward learning, and active inference. *Front. Comput. Neurosci.* 9:136. doi: 10.3389/fncom.2015.00136

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput. Biol.* 4:e1000211. doi: 10.1371/journal.pcbi.1000211

Friston, K. (2011). What is optimal about motor control? *Neuron* 72, 488–498. doi: 10.1016/j.neuron.2011.10.018

Friston, K., Adams, R. A., Perrinet, L., and Breakspear, M. (2012a). Perceptions as hypotheses: saccades as experiments. *Front. Psychol.* 3:151. doi: 10.3389/fpsyg.2012.00151

Friston, K., and Buzsáki, G. (2016). The functional anatomy of time: what and when in the brain. *Trends Cogn. Sci.* 20, 500–511. doi: 10.1016/j.tics.2016.05.001

Friston, K., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS ONE* 4:e6421. doi: 10.1371/journal.pone.0006421

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017a). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Friston, K., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017b). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Friston, K., Parr, T., and de Vries, B. (2017c). The graphical brain: belief propagation and active inference. *Network Neurosci.* 1, 381–414. doi: 10.1162/NETN_a_00018

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–214. doi: 10.1080/17588928.2015.1020053

Friston, K., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017d). Deep temporal models and active inference. *Neurosci. Biobehav. Rev.* 77, 388–402. doi: 10.1016/j.neubiorev.2017.04.009

Friston, K., Samothrakis, S., and Montague, R. (2012b). Active inference and agency: optimal control without cost functions. *Biol. Cybernet.* 106, 523–541. doi: 10.1007/s00422-012-0512-8

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Front. Hum. Neurosci.* 7:598. doi: 10.3389/fnhum.2013.00598

Girshick, A. R., Landy, M. S., and Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat. Neurosci.* 14:926. doi: 10.1038/nn.2831

Gottlieb, J., and Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nat. Rev. Neurosci.* 19, 758–770. doi: 10.1038/s41583-018-0078-0

Hassabis, D., and Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends Cogn. Sci.* 11, 299–306. doi: 10.1016/j.tics.2007.05.001

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* 28, 2539–2550. doi: 10.1523/JNEUROSCI.5487-07.2008

Huang, Y., and Rao, R. P. (2011). Predictive coding. *Wiley Interdiscipl. Rev. Cogn. Sci.* 2, 580–593. doi: 10.1002/wcs.142

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Res.* 49, 1295–1306. doi: 10.1016/j.visres.2008.09.007

Jóhannesson, M. I., Thornton, I. M., Smith, I. J., Chetverikov, A., and Kristjánsson, r. (2016). Visual foraging with fingers and eye gaze. *I-Perception* 7:2041669516637279. doi: 10.1177/2041669516637279

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). "Empowerment: a universal agent-centric measure of control," in *2005 IEEE Congress on Evolutionary Computation*, Vol. 1 (Edinburgh: IEEE), 128–135. doi: 10.1109/CEC.2005.1554676

Körding, K. P., and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature* 427:244. doi: 10.1038/nature02169

Lee, T. S., and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* 20, 1434–1448. doi: 10.1364/JOSAA.20.001434

Linsker, R. (1990). Perceptual neural organization: Some approaches based on network models and information theory. *Annu. Rev. Neurosci.* 13, 257–281. doi: 10.1146/annurev.ne.13.030190.001353

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organ. Sci.* 2, 71–87. doi: 10.1287/orsc.2.1.71

Millidge, B., Tschantz, A., Seth, A. K., and Buckley, C. L. (2020). On the relationship between active inference and control as inference. *arXiv* 2006.12964.

Mirza, M. B., Adams, R. A., Friston, K., and Parr, T. (2019a). Introducing a bayesian model of selective attention based on active inference. *Sci. Rep.* 9, 1–22. doi: 10.1038/s41598-019-50138-8

Mirza, M. B., Adams, R. A., Mathys, C., and Friston, K. (2018). Human visual exploration reduces uncertainty about the sensed world. *PLoS ONE* 13:e0190429. doi: 10.1371/journal.pone.0190429

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. (2016). Scene construction, visual foraging, and active inference. *Front. Comput. Neurosci.* 10:56. doi: 10.3389/fncom.2016.00056

Mirza, M. B., Adams, R. A., Parr, T., and Friston, K. (2019b). Impulsivity and active inference. *J. Cogn. Neurosci.* 31, 202–220. doi: 10.1162/jocn_a_01352

Narayanan, S., and Jurafsky, D. (1998). "Bayesian models of human sentence processing," in *Proceedings of the Twelfth Annual Meeting of the Cognitive Science Society* (Cambridge, MA), 1–6.

Ognibene, D., and Baldassare, G. (2014). Ecological active vision: four bioinspired principles to integrate bottom-up and adaptive top-down attention tested with a simple camera-arm robot. *IEEE Trans. Auton. Mental Dev.* 7, 3–25. doi: 10.1109/TAMD.2014.2341351

Ognibene, D., and Demiris, Y. (2013). "Towards active event recognition," in *Twenty-Third International Joint Conference on Artificial Intelligence* (Beijing).

Ólafsdóttir, I. M., Gestsdóttir, S., and Kristjánsson, A. (2019). Visual foraging and executive functions: a developmental perspective. *Acta Psychol.* 193, 203–213. doi: 10.1016/j.actpsy.2019.01.005

Palmer, J., Huk, A. C., and Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J. Vis.* 5:1. doi: 10.1167/5.5.1

Parr, T. (2020). Inferring what to do (and what not to). *Entropy* 22:536. doi: 10.3390/e22050536

Parr, T., Benrimoh, D. A., Vincent, P., and Friston, K. (2018). Precision and false perceptual inference. *Front. Integr. Neurosci.* 12:39. doi: 10.3389/fnint.2018.00039

Parr, T., and Friston, K. (2017). Uncertainty, epistemics and active inference. *J. R. Soc. Interface* 14:376. doi: 10.1098/rsif.2017.0376

Parr, T., and Friston, K. (2018a). The anatomy of inference: generative models and brain structure. *Front. Comput. Neurosci.* 12:90. doi: 10.3389/fncom.2018.00090

Parr, T., and Friston, K. (2018b). Attention or salience? *Curr. Opin. Psychol.* 29, 1–5. doi: 10.1016/j.copsyc.2018.10.006

Parr, T., and Friston, K. (2018c). The discrete and continuous brain: from decisions to movement-and back again. *Neural Comput.* 30, 2319–2347. doi: 10.1162/neco_a_01102

Parr, T., and Friston, K. (2019). Generalised free energy and active inference. *Biol. Cybernet.* 113, 495–513. doi: 10.1007/s00422-019-00805-w

Parr, T., Markovic, D., Kiebel, S. J., and Friston, K. (2019). Neuronal message passing using mean-field, bethe, and marginal approximations. *Sci. Rep.* 9:1889. doi: 10.1038/s41598-018-38246-3

Pezzulo, G., Rigoli, F., and Friston, K. (2018). Hierarchical active inference: a theory of motivated control. *Trends Cogn. Sci.* 22, 294–306. doi: 10.1016/j.tics.2018.01.009

Pineau, J., Roy, N., and Thrun, S. (2001). "A hierarchical approach to POMDP planning and execution," in *ICML Workshop on Hierarchy and Memory in Reinforcement Learning* (Williamstown, MA).

Puterman, M. L. (1995). Markov decision processes: discrete stochastic dynamic programming. *J. Oper. Res. Soc.* 46, 792–792. doi: 10.2307/2584317

Quétard, B., Quinton, J. C., Mermillod, M., Barca, L., Pezzulo, G., Colomb, M., et al. (2016). Differential effects of visual uncertainty and contextual guidance on perceptual decisions: evidence from eye and mouse tracking in visual search. *J. Vis.* 16, 28–28. doi: 10.1167/16.11.28

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev.* 85:59. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput.* 20, 873–922. doi: 10.1162/neco.2008.12-06-420

Rayner, K., and Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: a further examination. *Psychon. Bull. Rev.* 3, 504–509. doi: 10.3758/BF03214555

Runyan, C. A., Piasini, E., Panzeri, S., and Harvey, C. D. (2017). Distinct timescales of population coding across cortex. *Nature* 548:92. doi: 10.1038/nature23020

Schmidhuber, J. (1991). "Curious model-building control systems," in *Proceedings of International Joint Conference on Neural Networks* (Singapore), 1458–1463. doi: 10.1109/IJCNN.1991.170605

Seth, A. K. (2015). *The Cybernetic Bayesian Brain: From Interoceptive Inference to Sensorimotor Contingencies: From Interoceptive Inference to Sensorimotor Contingencies*. Sussex: Open MIND.

Seth, A. K., and Tsakiris, M. (2018). Being a beast machine: the somatic basis of selfhood. *Trends Cogn. Sci.* 22, 969–981. doi: 10.1016/j.tics.2018.08.008

Shadlen, M. N., and Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proc. Natl. Acad. Sci. U.S.A.* 93, 628–633. doi: 10.1073/pnas.93.2.628

Sridharan, M., Wyatt, J., and Dearden, R. (2010). Planning to see: a hierarchical approach to planning visual actions on a robot using POMDPs. *Artif. Intell.* 174, 704–725. doi: 10.1016/j.artint.2010.04.022

Stocker, A. A., and Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* 9:578. doi: 10.1038/nn1669

Sutton, R., and Barto, A. (1998). Reinforcement learning: an introduction. Cambridge, MA: MIT Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Theocharous, G., Murphy, K., and Kaelbling, L. P. (2004). "Representing hierarchical POMDPs as DBNS for multi-scale robot localization," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04*, Vol. 1 (New Orleans, LA: IEEE), 1045–1051. doi: 10.1109/ROBOT.2004.1307288

Todorov, E. (2008). "General duality between optimal control and estimation," in *2008 47th IEEE Conference on Decision and Control* (Cancun), 4286–4292. doi: 10.1109/CDC.2008.4739438

Trueswell, J. C., Tanenhaus, M. K., and Garnsey, S. M. (1994). Semantic influences on parsing: use of thematic role information in syntactic ambiguity resolution. *J. Mem. Lang.* 33, 285–318. doi: 10.1006/jmla.1994.1014

Ungerleider, L. G., and Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* 4, 157–165. doi: 10.1016/0959-4388(94)90066-3

van den Broek, L., Wiegerinck, W., and Kappen, H. J. (2010). "Risk sensitive path integral control," in *26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)* (Catalina Island, CA: AUAI Press).

Wald, A., and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Stat.* 19, 326–339. doi: 10.1214/aoms/1177730197

Yang, S. C.-H., Lengyel, M., and Wolpert, D. M. (2016). Active sensing in the categorization of visual patterns. *eLife* 5:e12215. doi: 10.7554/eLife.12215

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York, NY: Plenum Press.

Zeidman, P., Lutti, A., and Maguire, E. A. (2015). Investigating the functions of subregions within anterior hippocampus. *Cortex* 73, 240–256. doi: 10.1016/j.cortex.2015.09.002

Zeki, S., Goodenough, O., and Zak, P. J. (2004). Neuroeconomics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 359, 1737–1748. doi: 10.1098/rstb.2004.1544

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

We provide the derivation of Equation (8), the expected free energy as an upper bound on the negative information gain and negative extrinsic value:

$$\begin{aligned}
\mathbf{G}(\tau,\pi) &= \mathbb{E}_{Q(o_\tau,s_\tau|\pi)}[\ln Q(s_\tau|\pi) - \ln P(o_\tau,s_\tau)] \\
&= \mathbb{E}_{Q(o_\tau,s_\tau|\pi)}[\ln Q(s_\tau|\pi) - \ln P(o_\tau,s_\tau) \\
&+ \underbrace{\ln Q(s_\tau|o_\tau,\pi) - \ln Q(s_\tau|o_\tau,\pi)}_{=0}] \\
&= \mathbb{E}_{Q(o_\tau,s_\tau|\pi)}[\ln Q(s_\tau|\pi) - \ln Q(s_\tau|o_\tau,\pi) \\
&- \ln P(o_\tau)] + \underbrace{\mathbb{E}_{Q(o_\tau|\pi)}[\mathbb{E}_{Q(s_\tau|o_\tau,\pi)}[\ln \frac{Q(s_\tau|o_\tau,\pi)}{P(s_\tau|o_\tau)}]]}_{\textit{expected KL divergence} \geq 0} \\
&\geq \mathbb{E}_{Q(o_\tau,s_\tau|\pi)}[\ln Q(s_\tau|\pi) - \ln Q(s_\tau|o_\tau,\pi) \\
&- \ln P(o_\tau)] \\
\implies \mathbf{G}(\tau,\pi) &\geq -\mathbb{E}_{Q(o_\tau|\pi)}[\mathrm{D}_{KL}[Q(s_\tau|o_\tau,\pi)||Q(s_\tau|\pi)]] \\
&- \mathbb{E}_{Q(o_\tau|\pi)}[\ln P(o_\tau)] \qquad \text{(i)}
\end{aligned}$$

We also offer a derivation of Equation (9), the formulation of the expected free energy as the sum of "risk" and "ambiguity," starting from its definition as an upper bound on the (negative) epistemic and instrumental values. We can write $\mathbf{G}$ for a given future time point τ and policy π as follows:

$$\mathbf{G}(\tau,\pi) \geq \underbrace{-\mathrm{E}_{Q(o_\tau|\pi)}[\mathrm{D}_{KL}[Q(s_\tau|o_\tau,\pi)\|Q(s_\tau|\pi)]]}_{\textit{Epistemic value}}$$

$$\begin{aligned}
&\underbrace{-\mathrm{E}_{Q(o_\tau|\pi)}[\ln P(o_\tau)]}_{\textit{Instrumental value}} \\
&= -\mathbb{E}_{Q(o_\tau|\pi)}[D_{KL}[Q(s_\tau|o_\tau,\pi)||Q(s_\tau|\pi)] \\
&+ \underbrace{\ln Q(o_\tau|\pi) - \ln Q(o_\tau|\pi)}_{=0}] - \mathrm{E}_{Q(o_\tau|\pi)}[\ln P(o_\tau)] \\
&= -\mathbb{E}_{Q(o_\tau|\pi)}[\mathbb{E}_{Q(s_\tau|o_\tau,\pi)}[\ln \frac{Q(s_\tau|o_\tau,\pi)Q(o_\tau|\pi)}{Q(s_\tau|\pi)Q(o_\tau|\pi)}]] \\
&- \mathbb{E}_{Q(o_\tau|\pi)}[\ln P(o_\tau)] \\
&= -\mathbb{E}_{Q(s_\tau|\pi)P(o_\tau|s_\tau)}[\ln \frac{Q(s_\tau|\pi)P(o_\tau|s_\tau)}{Q(s_\tau|\pi)Q(o_\tau|\pi)}] \\
&- \mathbb{E}_{Q(o_\tau|\pi)}[\ln P(o_\tau)] \\
&= \underbrace{\mathbb{E}_{Q(s_\tau|\pi)}\left[\mathrm{H}[P(o_\tau|s_\tau)]\right]}_{\textit{Ambiguity}} + \underbrace{D_{KL}[Q(o_\tau|\pi)||P(o_\tau)]}_{\textit{Risk}} \qquad \text{(ii)}
\end{aligned}$$

The above derivation assumes that the mapping from predicted states $Q(s_\tau|\pi)$ to predicted observations $Q(o_\tau|s_\tau,\pi)$ is given as the likelihood of the generative model, i.e., $Q(o_\tau,s_\tau|\pi) = P(o_\tau|s_\tau)Q(s_\tau|\pi)$.

We provide a derivation of Equation (10), the full variational free energy of the posterior over observations, hidden states and policies:

$$\begin{aligned}
F &= \mathrm{E}_{Q(\tilde{s},\pi)}[\ln Q(\tilde{s},\pi) - \ln P(\tilde{o},\tilde{s},\pi)] \\
&= -\mathrm{E}_{Q(\tilde{s},\pi)}[\ln P(\tilde{o},\tilde{s},\pi)] - \mathrm{H}[Q(\tilde{s},\pi)] \\
&= \mathrm{E}_{Q(\pi)}\left[-\mathrm{E}_{Q(\tilde{s}|\pi)}[\ln P(\tilde{o},\tilde{s}|\pi)] - \mathrm{H}[Q(\tilde{s}|\pi)]\right] \\
&+ \mathrm{D}_{KL}[Q(\pi)||P(\pi)] \\
&= \mathrm{E}_{Q(\pi)}[F(\pi)] + \mathrm{D}_{KL}[Q(\pi)||P(\pi)] \qquad \text{(iii)}
\end{aligned}$$

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 09 April 2021
doi: 10.3389/frai.2021.531316

An Overcomplete Approach to Fitting Drift-Diffusion Decision Models to Trial-By-Trial Data

Q. Feltgen[1] *and J. Daunizeau*[1,2]*

[1]Paris Brain Institute (ICM), Sorbonne Université, Inserm, CNRS, Hôpital Pitié-Salpêtrière, Paris, France, [2]ETH, Zurich, Switzerland

Drift-diffusion models or DDMs are becoming a standard in the field of computational neuroscience. They extend models from signal detection theory by proposing a simple mechanistic explanation for the observed relationship between decision outcomes and reaction times (RT). In brief, they assume that decisions are triggered once the accumulated evidence in favor of a particular alternative option has reached a predefined threshold. Fitting a DDM to empirical data then allows one to interpret observed group or condition differences in terms of a change in the underlying model parameters. However, current approaches only yield reliable parameter estimates in specific situations (c.f. fixed drift rates vs drift rates varying over trials). In addition, they become computationally unfeasible when more general DDM variants are considered (e.g., with collapsing bounds). In this note, we propose a fast and efficient approach to parameter estimation that relies on fitting a "self-consistency" equation that RT fulfill under the DDM. This effectively bypasses the computational bottleneck of standard DDM parameter estimation approaches, at the cost of estimating the trial-specific neural noise variables that perturb the underlying evidence accumulation process. For the purpose of behavioral data analysis, these act as nuisance variables and render the model "overcomplete," which is finessed using a variational Bayesian system identification scheme. However, for the purpose of neural data analysis, estimates of neural noise perturbation terms are a desirable (and unique) feature of the approach. Using numerical simulations, we show that this "overcomplete" approach matches the performance of current parameter estimation approaches for simple DDM variants, and outperforms them for more complex DDM variants. Finally, we demonstrate the added-value of the approach, when applied to a recent value-based decision making experiment.

Keywords: DDM, decision making, computational modeling, variational bayes, neural noise

OPEN ACCESS

Edited by:
Thomas Parr,
University College London,
United Kingdom

Reviewed by:
Sebastian Gluth,
University of Hamburg, Germany
Vincent Moens,
Catholic University of Louvain,
Belgium

***Correspondence:**
J. Daunizeau
jean.daunizeau@gmail.com

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

Received: *31 January 2020*
Accepted: *17 February 2021*
Published: *09 April 2021*

Citation:
Feltgen Q and Daunizeau J (2021) An Overcomplete Approach to Fitting Drift-Diffusion Decision Models to Trial-By-Trial Data. Front. Artif. Intell. 4:531316. doi: 10.3389/frai.2021.531316

INTRODUCTION

Over the past two decades, neurocognitive processes of decision making have been extensively studied under the framework of so-called *drift-diffusion models* or DDMs. These models tie together decision outcomes and response times (RT) by assuming that decisions are triggered once the accumulated evidence in favor of a particular alternative option has reached a predefined threshold (Ratcliff and McKoon, 2008; Ratcliff et al., 2016). They owe their popularity both to experimental successes in capturing observed data in a broad set of behavioral studies (Gold and Shadlen, 2007; Resulaj et al., 2009; Milosavljevic et al., 2010; De Martino et al., 2012; Hanks et al., 2014; Pedersen

et al., 2017), and to theoretical work showing that DDMs can be thought of as optimal decision problem solvers (Bogacz et al., 2006; Balci et al., 2011; Drugowitsch et al., 2012; Zhang, 2012; Tajima et al., 2016). Critically, mathematical analyses of the DDM soon demonstrated that it suffers from inherent non-identifiability issues, e.g., predicted choices and RTs are invariant under any arbitrary rescaling of DDM parameters (Ratcliff and Tuerlinckx, 2002; Ratcliff et al., 2016). This is important because, in principle, this precludes proper, quantitative, DDM-based data analysis. Nevertheless, over the past decade, many statistical approaches to DDM parameter estimation have been proposed, which yield efficient parameter estimation under simplifying assumptions (Voss and Voss, 2007; Wagenmakers et al., 2007, 2008; Vandekerckhove and Tuerlinckx, 2008; Grasman et al., 2009; Zhang, 2012; Wiecki et al., 2013; Zhang et al., 2014; Hawkins et al., 2015; Voskuilen et al., 2016; Pedersen and Frank, 2020). Typically, these techniques essentially fit the choice-conditional distribution of observed RT (or moments thereof), having arbitrarily fixed at least one of the DDM parameters. They are now established statistical tools for experimental designs where the observed RT variability is mostly induced by internal (e.g., neural) stochasticity in the decision process (Boehm et al., 2018).

Now current decision making experiments typically consider situations in which decision-relevant variables are manipulated on a trial-by-trial basis. For example, the reliability of perceptual evidence (e.g., the psychophysical contrast in a perceptual decision) may be systematically varied from one trial to the next. Under current applications of the DDM, this implies that some internal model variables (e.g., the drift rate) effectively vary over trials. Classical DDM parameter estimation approaches do not optimally handle this kind of experimental designs, because these lack the trial repetitions that would be necessary to provide empirical estimates of RT moments in each condition. In turn, alternative statistical approaches to parameter estimation have been proposed, which can exploit predictable inter-trial variations of DDM variables to fit the model to RT data (Wabersich and Vandekerckhove, 2014; Moens and Zenon, 2017; Pedersen et al., 2017; Fontanesi et al., 2019a; Fontanesi et al., 2019b; Gluth and Meiran, 2019). In brief, they directly compare raw RT data with expected RTs, which vary over trials in response to known variations in internal variables. Although close to optimal from a statistical perspective, they suffer from a challenging computational bottleneck that lies in the trial-by-trial derivation of RT first-order moments. This is why they are typically constrained to simple DDM variants, for which analytical solutions exist (Navarro and Fuss, 2009; Srivastava et al., 2016; Fengler et al., 2020; Shinn et al., 2020).

This note is concerned with the issue of obtaining reliable DDM parameter estimates from concurrent trial-by-trial choice and response time data, for a broad class of DDM variants. We propose a fast and efficient approach that relies on fitting a *self-consistency* equation, which RTs necessarily fulfill under the DDM. This provides a simple and elegant way to bypassing the common computational bottleneck of existing approaches, at the cost of considering additional trial-specific nuisance model variables. These are the cumulated "neural" noise that perturbs the evidence accumulation process at the corresponding trial. Including these variables in the model makes it "overcomplete," the identification of which is finessed with a dedicated variational Bayesian scheme. In turn, the ensuing overcomplete approach generalizes to a large class of DDM model variants, without any additional computational and/or implementational burden.

In *Model Formulation and Impact of DDM Parameters* section of this document, we briefly recall the derivation of the DDM, and summarize the impact of DDM parameters onto the conditional RT distributions. In *A Self-Consistency Equation for DDMs* and An *Overcomplete Likelihood Approach to DDM Inversion* sections, we derive the DDM's self-consistency equation and describe the ensuing overcomplete approach to DDM-based data analysis. In *Parameter Recovery Analysis* section, we perform parameter recovery analyses for standard DDM fitting procedures and the overcomplete approach. In *Application to a Value-Based Decision Making Experiment* section, we demonstrate the added-value of the overcomplete approach, when applied to a value-based decision making experiment. Finally, in *Discussion* section, we discuss our results in the context of the existing literature. In particular, we comment on the potential utility of neural noise perturbation estimates for concurrent neuroimaging data analysis.

MODEL FORMULATION AND IMPACT OF DDM PARAMETERS

First, let us recall the simplest form of a drift-diffusion decision model or DDM (in what follows, we will refer to this variant as the "vanilla" DDM). Let $x(t)$ be a decision variable that captures the accumulated evidence (up to time t) in favor of a given option in a binary choice set. Under the vanilla DDM, a decision is triggered whenever $x(t)$ hits either of two bounds, which are positioned at $x = b$ and $x = -b$, respectively. When a bound hit occurs defines the decision time, and which bound is hit determines the (binary) decision outcome o. By assumption, the decision variable $x(t)$ is supposed to follow the following stochastic differential equation:

$$dx = \breve{v} \times dt + \breve{\sigma} \times d\eta \tag{1}$$

where v is the drift rate, $d\eta \sim N(0, dt)$ is a standard Wiener process, and $\breve{\sigma}$ is the standard deviation of the stochastic (diffusion) perturbation term.

Equation 1 can be discretized using an Euler-Maruyma scheme (Kloeden and Platen, 1992), yielding the following discrete form of the decision variable dynamics:

$$x_{t+1} = x_t + v + \sigma\eta_t \tag{2}$$

where t indexes time on a temporal grid with resolution Δt, $v = \breve{v}\Delta t$ is the discrete-time drift rate, $\sigma = \breve{\sigma}\sqrt{\Delta t}$ is the discrete-time standard deviation of the perturbation term and $\eta_t \sim N(0, 1)$ is a standard normal random variable. By convention, the system's initial condition is denoted as x_0, which we refer to as the "initial bias".

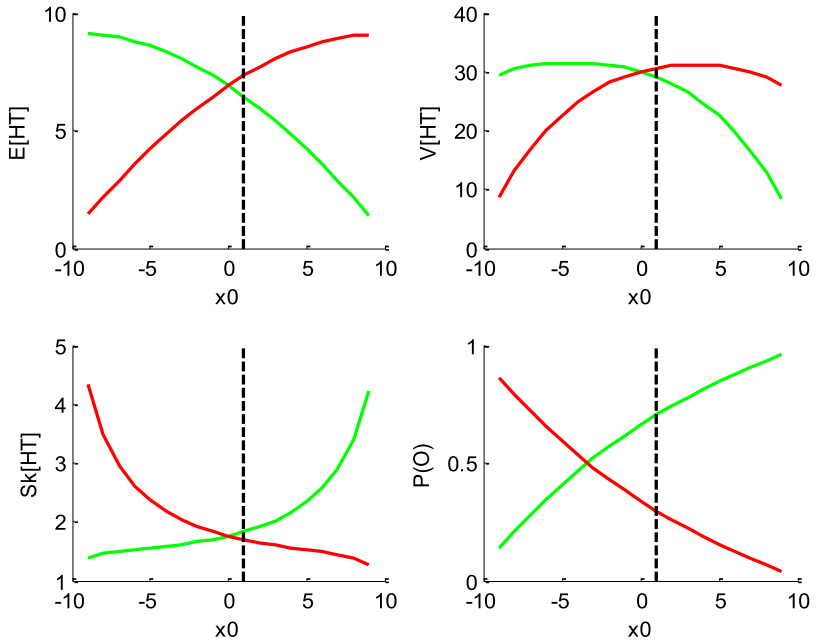


FIGURE 1 | Impact of initial bias x_0. In all panels, the color code indicates the decision outcomes (green: "up" decisions, red: "down" decisions). The black dotted line indicates the default parameter value (for ease of comparison with other figures below). Upper-left panel: mean hitting times (y-axis) as a function of initial bias (x-axis). Upper-right panel: hitting times' variance (y-axis) as a function of initial bias (x-axis). Lower-left panel: hitting times' skewness (y-axis) as a function of initial bias (x-axis). Lower-right panel: outcome ratios (y-axis) as a function of initial bias (x-axis).

The joint distribution of response times and decision outcomes depends upon the DDM parameters, which include: the drift rate v, the bound's height b, the noise's standard deviation σ and the initial condition x_0. DDMs also typically include a so-called "non-decision" time parameter T_{ND}, which captures systematic latencies between covert bound hit times and overt response times. Under such simple DDM variant, variability in response times and decision outcomes derive from stochastic terms η. These are typically thought of as neural noise that perturb the evidence accumulation process within the brain's decision system (Gold and Shadlen, 2007; Turner et al., 2015; Guevara Erra et al., 2019).

Under such simple DDM variant, analytical expressions exist for the first two moments of RT distributions (Srivastava et al., 2016). Higher-order moments can also be derived from efficient semi-analytical solutions to the issue of deriving the joint choice/RT distribution (Navarro and Fuss, 2009). However, more complex variants of the DDM (including, e.g., collapsing bounds) are much more difficult to simulate, and require either sampling schemes or numerical solvers of the underlying Fokker-Planck equation (Fengler et al., 2020; Shinn et al., 2020).

Figures 1–4 below demonstrate the impact of model parameters on the decision outcome ratios $P(o|v, x_0, b, \sigma)$ and the first three moments of conditional hitting time (HT) distributions, namely: their mean $E[HT|o, v, x_0, b, \sigma]$, variance $V[HT|o, v, x_0, b, \sigma]$ and skewness $S_k[HT|o, v, x_0, b, \sigma]$. As we will see, each DDM parameter has a specific signature, in terms of its joint impact on these seven quantities. This does not imply however, that different parameter settings necessarily yield distinct moments. In fact, there are changes in the DDM parameters that leave the predicted moments unchanged. This will induce parameter recovery issues, which we will demonstrate later.

But let first us summarize the impact of DDM parameters. To do this, we first set model parameters to the following "default" values: $v = 1/2$, $x_0 = 1$, $b = 10$ and $\sigma = 4$. This parameter setting yields about 30% decision errors, which we take as a valid reference point for typical studies of decision making. In what follows, we vary each model parameter one by one, keeping the other ones at their default value.

Figure 1 below shows the impact of initial bias x_0.

One can see that increasing the initial bias accelerates decision times for "up" decisions, and decelerates decision times for "down" decisions. This is because increasing x_0 mechanically increases the probability of an early upper bound hit, and decreases the probability of an early lower bound hit. Increasing x_0 also decreases (resp., increases) the variance for "up" (resp., "down") decisions, and increases (resp., decreases) the skewness for "up" (resp., "down") decisions. Finally, increasing the initial bias increases the ratio of "up" decisions. These are corollary consequences of increasing (resp. decreasing) the probability of an early upper (resp., lower) bound hit. This is because when an increasing proportion of stochastic paths eventually hit a bound very early, this squeezes the distribution of hitting times just above null hitting times. Note that the outcome ratios are not equal to 1/2 when $x_0 = 0$. This is because the default drift rate v is positive, and therefore favors

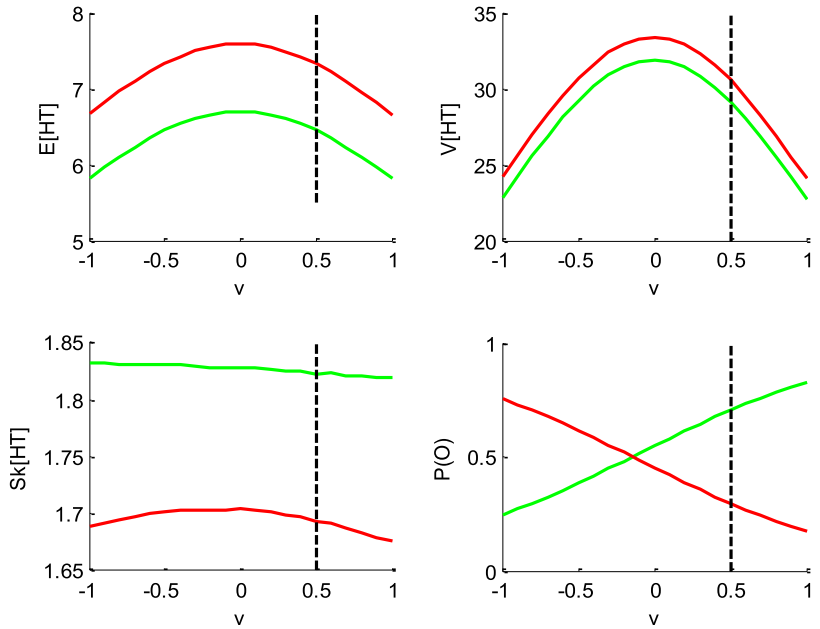


FIGURE 2 | Impact of drift rate v. Same format as **Figure 1**.

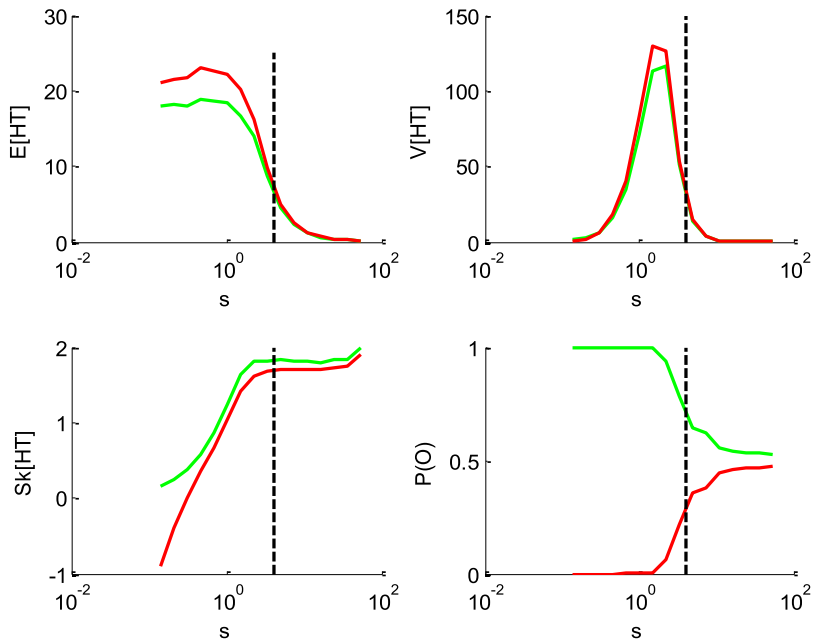


FIGURE 3 | Impact of the perturbation' standard deviation σ. Same format as **Figure 1** (but the x-axis is now in log-scale).

"up" decisions. Most importantly, the initial bias is the only DDM parameter that has opposite effects on mean HT for "up" and "down" decision outcomes.

Figure 2 below shows the impact of drift rate v.

One can see that the mean and variance of decision times are maximal when the drift rate is null. This is because the probability of an early (upper or lower) bound hit decreases as $v \to 0$. Also, the drift rate has little impact on the HT skewness. Note that, in contrast to the initial bias, the impact of the drift rate on mean HT is identical for both "up" and "down" decisions. Finally, and as expected, increasing the drift rate increases the ratio of "up" decisions.

Figure 3 below shows the impact of the noise's standard deviation σ.

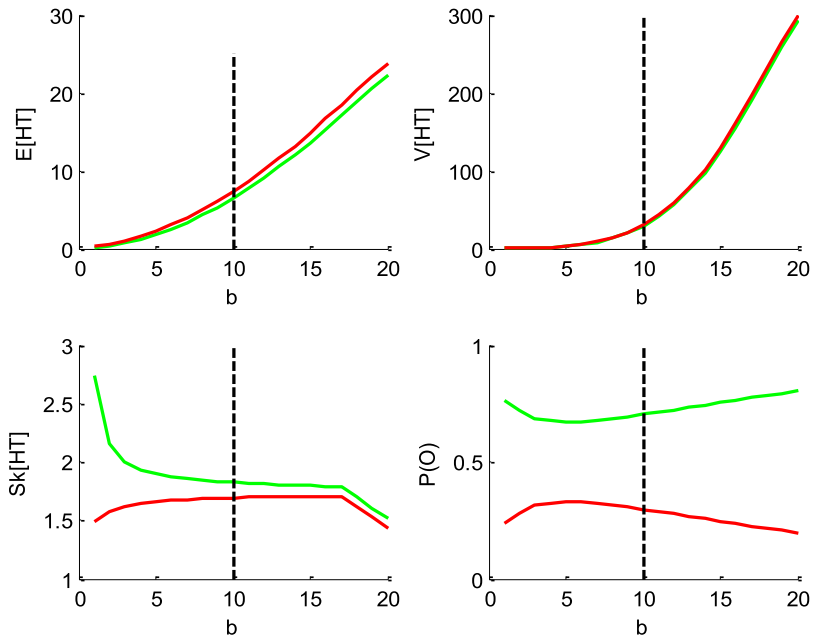


FIGURE 4 | Impact of the threshold's height b. Same format as **Figure 1**.

One can see that increasing the standard deviation decreases the mean HT, and increases its skewness. This is, again, because increasing σ increases the probability of an early bound hit. Its impact on the variance, however, is less trivial. When the standard deviation σ is very low, increasing σ first increases the hitting times' variance. This is because it progressively frees the system from its deterministic fate, therefore enabling HT variability around the mean. Then, it reaches a critical point above which increasing it further now decreases the variance. This is again a consequence of increasing the probability of an early bound hit. The associated HT squeezing effect can be seen on the skewness, which steadily increases beyond the critical point. Note that the standard deviation has the same impact on mean HT for "up" and "down" decisions. Finally, increasing the standard deviation eventually maximizes the entropy of the decision outcomes, i.e., $P(o) \to 1/2$ when $\sigma \to \infty$. This is because the relative contribution of the diffusion term eventually masks the drift.

Figure 4 below shows the impact of the bound's height b.

One can see that increasing the bound's height increases both the mean and the variance of HT, and decreases its skewness, identically for "up" and "down" decisions. Finally, increasing the threshold's height decreases the entropy of the decision outcomes, i.e., $P(o) \to 0$ or 1 when $b \to \infty$. This directly derives from the fact that increasing b decreases the probability of an early bound hit. This effect basically competes with the effect of the standard deviation σ, which accelerates HTs. This is why one may say that increasing the threshold's height effectively increases the demand for evidence strength in favor of one of the decision outcomes.

Note that the impact of the "non-decision" time T_{ND} simply reduces to shifting the mean of the RT distribution, without any effect on other moments.

In brief, DDM parameters have distinct impacts on the sufficient statistics of response times. This means that they could, in principle, be discriminated from each other. Standard DDM fitting procedures rely on adjusting the DDM parameters so that the RT moments (e.g., up to third order) match model predictions. In what follows, we refer to this as the "method of moments" (see **Supplementary Appendix S2**). However, we will see below that the DDM is not perfectly identifiable. One can also see that changing any of these parameters from trial to trial will most likely induce non-trivial variations in RT data. Here, the method of moments may not be optimal, because predictable trial-by-trial variations in DDM parameters will be lumped together with stochastic perturbation-induced variations. One may thus rather attempt to match the trial-by-trial series of raw response times directly with their corresponding first-order moments. In what follows, we refer to this as the "method of trial means" (see **Supplementary Appendix S3**). Given the computational cost of deriving expected response times for each trial, this type of approach is typically restricted to the vanilla DDM, since there is no known analytical expression for response time moments under more complex DDM variants.

Below, we suggest a simple and efficient way of performing DDM parameter estimation, which applies to a broad class of DDM variants without significant additional computational burden. This follows from fitting a self-consistency equation that, under a broad class of DDM variants, response times have to obey.

A SELF-CONSISTENCY EQUATION FOR DDMS

First, note that **Eq. 2** can be rewritten as follows:

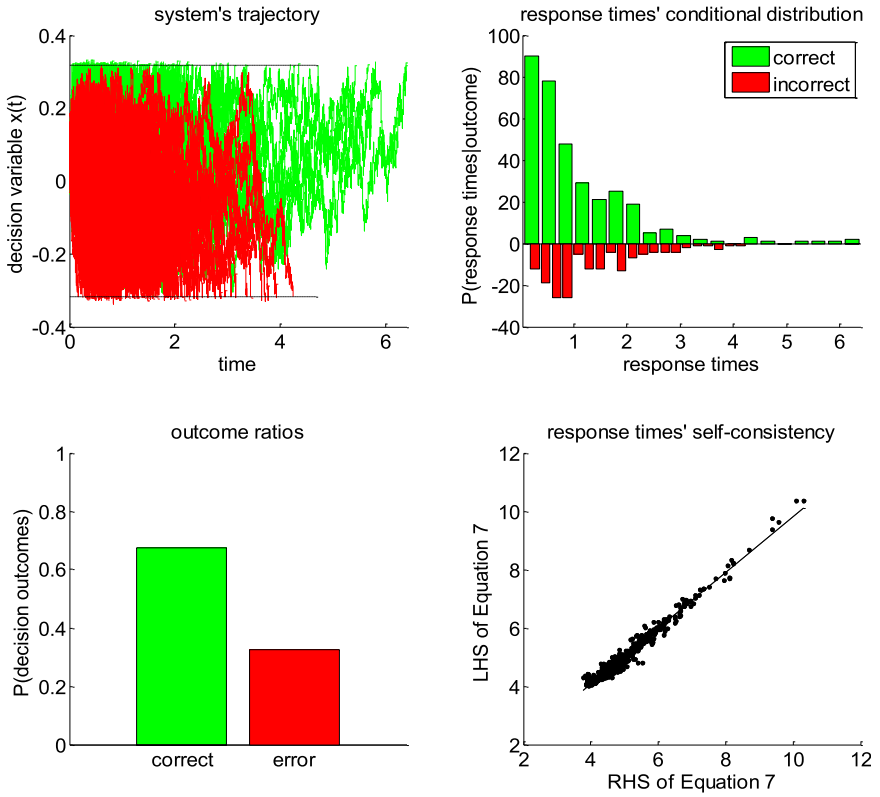


FIGURE 5 | Self-consistency equation. Monte-Carlo simulation of 200 trials of a DDM, with arbitrary parameters (in this example, the drift rate is positive). In all panels, the color code indicates the decision outcomes, which depends upon the sign of the drift rate (green: correct decisions, red: incorrect decisions). Upper-left panel: simulated trajectories of the decision variable (*y*-axis) as a function of time (*x*-axis). Upper-right panel: response times' distribution for both correct and incorrect choice outcomes over the 200 Monte-Carlo simulations. Lower-left panel: outcome ratios. Lower-right panel: the left-hand side of **Eq. 7** (*y*-axis) is plotted against the right-hand side of **Eq. 7** (*x*-axis), for each of the 200 trials.

$$\begin{aligned} x_t &= x_0 + tv + \sigma \sum_{t'=0}^{t-1} \eta_{t'} \\ &= x_0 + tv + \sigma\sqrt{t}\,\tilde{\eta}_t \end{aligned} \tag{3}$$

where we coin $\tilde{\eta}_t \triangleq 1/\sqrt{t}\sum_{t'=0}^{t-1}\eta_{t'}$ the "normalized cumulative perturbation". Now let τ_i be the decision time of the ith trial. Note that decision times are trivially related to cumulative perturbations because, by definition, $|x_{\tau_i}| = b$. This implies that:

$$b = \left| x_0 + \tau_i v + \sigma\sqrt{\tau_i}\tilde{\eta}_{\tau_i} \right| \tag{4}$$

where $\tilde{\eta}_{\tau_i}$ denotes the (unknown) cumulative perturbation term of the ith trial.

Information regarding the binary decision outcome $o_i \in \{-1, 1\}$ further disambiguates **Eq. 4** as follows:

$$\begin{aligned} b &= \begin{cases} x_0 + \tau_i v + \sigma\sqrt{\tau_i}\tilde{\eta}_{\tau_i} & \text{if } o_i = 1 \quad (\text{'up' decision}) \\ -x_0 - \tau_i v - \sigma\sqrt{\tau_i}\tilde{\eta}_{\tau_i} & \text{if } o_i = -1 \ (\text{'down' decision}) \end{cases} \\ &= o_i\left(x_0 + \tau_i v + \sigma\sqrt{\tau_i}\,\tilde{\eta}_{\tau_i}\right) \end{aligned} \tag{5}$$

where o_i can only take two possible values (−1 or 1). **Eq. 5** can then be used to relate decision times directly to DDM model parameters (and cumulative perturbations):

$$\tau_i = \frac{o_i b - x_0}{v} - \frac{\sigma\sqrt{\tau_i}}{v}\tilde{\eta}_{\tau_i} \tag{6}$$

From **Eq. 6**, one can express observed trial-by-trial empirical response times y_i as follows:

$$y_i \approx \frac{o_i b - x_0}{v} - \frac{\sigma\sqrt{y_i - T_{ND}}}{v}\tilde{\eta}_{\tau_i} + T_{ND} + \varepsilon_i \tag{7}$$

where ε_i are unknown i. i.d. model residuals.

Note that decision times effectively appear on both the left-hand and the right-hand side of **Eqs 6, 7**. This is a slightly unorthodox feature, but, as we will see, this has effectively no consequence from the perspective of model inversion. In fact, one can think of **Eq. 7** as a "self-consistency" constraint that response times have to fulfill under the DDM. This is why we refer to **Eq. 7** as the *self-consistency equation* of DDMs. This, however, prevents us from using **Eq. 7** to generate data under the DDM. In other terms, **Eq. 7** is only useful when analyzing empirical RT data.

Figure 5 below exemplifies the accuracy of DDM's self-consistency equation, using a Monte-Carlo simulation of 200 trials under the vanilla DDM.

One can see that the DDM's self-consistency equation is valid, i.e., simulated response times almost always equate their

theoretical prediction. The few (small) deviations that can be eyeballed on the lower-right panel of **Figure 5** actually correspond to simulation artifacts where the decision variable exceeds the bound by some relatively small amount. This happen when the discretization step Δt (cf. **Eq. 2**) is too large when compared to the relative magnitude of the stochastic component of the system's dynamics. In effect, these artifactual errors grow when σ/v increases. Nevertheless, in principle, these and other errors would be absorbed in the model residuals ε_i of **Eq. 7**.

Now recall that recent extensions of vanilla DDMs include e.g., collapsing bounds (Hawkins et al., 2015; Voskuilen et al., 2016) and/or nonlinear transformations of the state-space (Tajima et al., 2016). As the astute reader may have already guessed, the self-consistency equation can be generalized to such DDM variants. Let us assume that **Eqs 2, 3** still hold, i.e., the decision process is still somehow based upon a gaussian random walk. However, we now assume that the decision is triggered when an arbitrary transformation $z : x \to z(x)$ of the base random walk x_t has reached a predefined threshold $\widehat{b}(t)$ that can vary over time (e.g., a collapsing bound). **Eq. 5** now becomes:

$$\widehat{b}(\tau_i) = o_i z\left(x_0 + \tau_i v + \sigma\sqrt{\tau_i}\,\tilde{\eta}_i\right) \tag{8}$$

If the transformation $z : x \to z(x)$ is invertible (i.e., if z^{-1} exists and is unique), then the self-consistency equation for reaction times y_i now generalizes as follows:

$$y_i \approx \underbrace{\frac{z^{-1}\left[o_i \widehat{b}(y_i - T_{ND})\right] - x_0}{v} - \frac{\sigma\sqrt{y_i - T_{ND}}}{v}\tilde{\eta}_i}_{g\left(v, x_0, \sigma, T_{ND}, \tilde{\eta}_i\right)} + T_{ND} + \varepsilon_i \tag{9}$$

where $g(v, x_0, \sigma, T_{ND}, \tilde{\eta}_i)$ is the "expected" (or rather, "self-consistent") response time at trial i, which depends nonlinearly on DDM parameters (and on response times). Note that one recovers the self-consistency equation of "vanilla" DDM (**Eq. 7**) when setting $z(x) = z^{-1}(x) = x$ and $\widehat{b}(t) = b \;\; \forall t$.

Importantly, inverting **Eq. 9** can be used to estimate parameters γ and ω that control the transformation $z_\gamma : x \overset{\gamma}{\to} z_\gamma(x)$ or the collapsing bounds $\widehat{b}_\omega : t \overset{\omega}{\to} \widehat{b}_\omega(t)$, respectively. We will see examples of this in the Results section below. This implies that the self-consistency equation can be used, in conjunction with adequate statistical parameter estimation approaches (see below), for estimating DDM parameters under many different variants of DDM, including those for which no analytical result exists for the response time distribution.

AN OVERCOMPLETE LIKELIHOOD APPROACH TO DDM INVERSION

Fitting **Eq. 9** to response time data reduces to finding the set of parameters that renders the DDM self-consistent. In doing so, normalized cumulative perturbation terms $\tilde{\eta}$ are treated as nuisance model parameters, but model parameters nonetheless. This means that there are more model parameters than there are data points. In other words, **Eq. 9** induces an "overcomplete" likelihood function $p(y|v, x_0, \sigma, \omega, \gamma, T_{ND}, \tilde{\eta}, \lambda)$:

$$\begin{aligned} p(y|v, x_0, \sigma, \omega, \gamma, T_{ND}, \tilde{\eta}, \lambda) &= \prod_{i=1}^{n} p(y_i|v, x_0, \sigma, \omega, \gamma, T_{ND}, \tilde{\eta}_i, \lambda) \\ &= \prod_{i=1}^{n} N\left(g(v, x_0, \sigma, \omega, \gamma, T_{ND}, \tilde{\eta}_i), \lambda\right) \end{aligned} \tag{10}$$

where λ is the variance of the model residuals ε_i of **Eq. 9**, $g(\cdot)$ is the "self-consistent" response time given in **Eq. 9**, and we have used the (convenient but slightly abusive) notation $\tilde{\eta}_i$ to reference cumulative perturbations w.r.t. to their corresponding trial index.

Dealing with such overcomplete likelihood function requires additional constraints on model parameters: this is easily done within a Bayesian framework. Therefore, we rely on the variational Laplace approach (Friston et al., 2007; Daunizeau, 2017), which was developed to perform approximate bayesian inference on nonlinear generative models (see **Supplementary Appendix S1** for mathematical details). In what follows, we propose a simple set of prior constraints that help regularizing the inference.

a. Prior moments of the cumulative perturbations: the "no barrier" approximation

Recall that, under the DDM framework, errors can only be due to the stochastic perturbation noise. More precisely, errors are due to those perturbations that are strong enough to deviate the system's trajectory and make it hit the "wrong" bound (e.g., the lower bound if the drift rate is positive). Let $Q_=$ be the proportion of correct responses. For example, if the drift rate is positive, then $Q_=$ corresponds to responses that hit the upper bound. Now let $\tilde{\eta}_=$ be the critical value of $\tilde{\eta}$ such that $P(\tilde{\eta} \geq \tilde{\eta}_=) = Q_=$ (see **Figure 6** below). Then, we know that errors correspond to those perturbations $\tilde{\eta}_i$ that are smaller than $\tilde{\eta}_=$. But what do we know about the distribution of perturbations? Importantly, if the DDM's stochastic evidence accumulation process had no decision bound, then the distribution of normalized cumulative perturbations would be invariant over time and such that $\tilde{\eta}_t \sim N(0,1) \;\; \forall t$. This, in fact, is the very reason why we introduced normalized cumulative perturbations in **Eq. 3**. Under this "no barrier" approximation, one can now derive the conditional expectations $\tilde{\mu}_=$ and $\tilde{\mu}_{\neq}$ of the perturbation $\tilde{\eta}_i$, given that the decision outcome o_i is correct or erroneous, respectively:

$$\begin{cases} \tilde{\mu}_= \triangleq E\left[\tilde{\eta}_i | o_i = 1\right] = E\left[\tilde{\eta}_i | \tilde{\eta}_i > \tilde{\eta}_=\right] = \dfrac{1}{(1 - Q_=)\sqrt{2\pi}} \exp\left(-\dfrac{1}{2}\tilde{\eta}_=^2\right) \\ \tilde{\mu}_{\neq} \triangleq E\left[\tilde{\eta}_i | o_i = -1\right] = E\left[\tilde{\eta}_i | \tilde{\eta}_i < \tilde{\eta}_=\right] = -\dfrac{1}{Q_=\sqrt{2\pi}} \exp\left(-\dfrac{1}{2}\tilde{\eta}_=^2\right) \end{cases} \tag{11}$$

Equation 11 is obtained from the known expression of first-order moments of a truncated normal density $N(0,1)$. Critically, **Eq. 11** does not depend upon DDM parameters. Of course, the

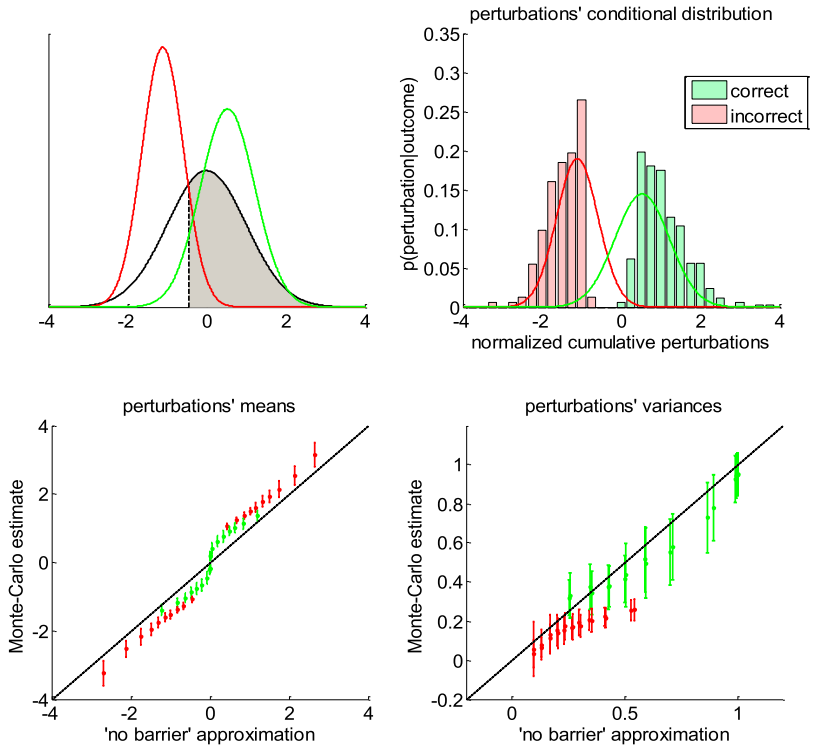


FIGURE 6 | Approximate conditional distributions of the normalized cumulative perturbations. Upper-left panel: The black line shows the "no barrier" standard normal distribution of normalized cumulative perturbations. The shaded gray area has size $Q_=$, and its left bound (dashed black line) is the critical value $\tilde{\eta}_=$ below which cumulative perturbations eventually induce errors. The green and red lines depict the ensuing approximate conditional distributions given in **Eq. 13**. Upper-right panel: a Representative monte-carlo simulation. The green and red bars show the sample histogram of normalized cumulative perturbations for correct and erroneous decisions, respectively (over 200 trials, same simulation as in **Figure 5**). The green and red lines depict the corresponding approximate conditional normal distributions (cf. **Eq. 13**). Lower-left panel: The sample mean estimates of conditional perturbations (*y*-axis) are plotted against their "no barrier" approximation (*x*-axis, **Eq. 11**). Monte-carlo simulations are split according to the sign of the drift rate, and then binned according to deciles of approximate conditional means of normalized cumulative perturbations (green: Correct, red: error, errorbars: Within-decile means ± standard deviations). The black dotted line shows the identity mapping (perfect approximation). Lower-right panel: The sample variance estimates of normalized cumulative perturbations (*y*-axis) are plotted against their "no barrier" approximation (*x*-axis, **Eq. 12**). Same format as lower-left panel.

same logic extends to conditional variances $\tilde{\Sigma}_=$ and $\tilde{\Sigma}_{\neq}$, whose analytical expressions are given by:

$$\begin{cases} \tilde{\Sigma}_= \triangleq V\left[\tilde{\eta}_i | o_i = 1\right] = V\left[\tilde{\eta}_i | \tilde{\eta}_i > \tilde{\eta}_=\right] = 1 + \tilde{\eta}_=\tilde{\mu}_= - \tilde{\mu}_=^2 \\ \tilde{\Sigma}_{\neq} \triangleq V\left[\tilde{\eta}_i | o_i = -1\right] = V\left[\tilde{\eta}_i | \tilde{\eta}_i < \tilde{\eta}_=\right] = 1 + \tilde{\eta}_=\tilde{\mu}_{\neq} - \tilde{\mu}_{\neq}^2 \end{cases} \quad (12)$$

A simple moment-matching approach thus suggests to approximate the conditional distribution $p(\tilde{\eta}_i | o_i)$ of normalized cumulative perturbations as follows:

$$p(\tilde{\eta}_i | o_i) = \begin{cases} N\left(\tilde{\mu}_=, \tilde{\Sigma}_=\right) & \text{if } o_i = correct \\ N\left(\tilde{\mu}_{\neq}, \tilde{\Sigma}_{\neq}\right) & \text{if } o_i = error \end{cases} \quad (13)$$

where the correct/error label depends on the sign of the drift rate. This concludes the derivation of our simple "no barrier" approximation to the conditional moments of cumulative perturbations.

Note that we derived this approximation without accounting for the (only) mathematical subtlety of the DDM: namely, the fact that decision bounds formally act as "absorbing barriers" for the system (Broderick et al., 2009). Critically, absorbing barriers induce some non-trivial forms of dynamical degeneracy. In particular, they eventually favor paths that are made of extreme samples of the perturbation noise. This is because these have a higher chance of crossing the boundary, despite being comparatively less likely than near-zero samples under the corresponding "no barrier" distribution. One may thus wonder whether ignoring absorbing barriers may invalidate the moment-matching approximation given in **Eqs 11–13**. To address this concern, we conducted a series of 1000 Monte-Carlo simulations, where DDM parameters were randomly drawn (each simulation consisted of 200 trials of the same decision system). We use these to compare the sample estimates of first- and second-order moments of normalized cumulative perturbations and their analytical approximations (as given in **Eqs. 11, 12**). The results are given in **Figure 6** below.

One can see on the upper-right panel of **Figure 6** that the distribution of normalized cumulative perturbations may strongly deviate from the standard normal density. In particular, this distribution clearly exhibits two modes, which correspond to correct and incorrect decisions, respectively. We have observed this bimodal shape across

almost all Monte-Carlo simulations. This means that bound hits are less likely to be caused by perturbations of small magnitude than expected under the "no-barrier" distribution (cf. lack of probability mass around zero). Nevertheless, the ensuing approximate conditional distributions seem to be reasonably matched with their sample estimates. In fact, lower panels of **Figure 6** demonstrate that sample means and variances of normalized cumulative perturbations are well approximated by **Eqs 11, 12** for a broad range of DDM parameters. We note that the "no-barrier" approximation tends to slightly underestimate first-order moments, and overestimate second-order moments. This bias is negligible however, when compared to the overall range of variations of conditional moments. In brief, the effect of absorbing barriers on system dynamics has little impact on the conditional moments of normalized cumulative perturbations.

When fitting the DDM to empirical RT data, one thus wants to enforce the distributional constraint in **Eqs 11–13** onto the perturbation term in **Eq. 9**. This can be done using a change of variable $\tilde{\eta}_i = h(\varsigma_i)$, where ς are non-constrained dummy variables and $h : \varsigma_i \rightarrow h(\varsigma_i)$ is the following moment-enforcing mapping:

$$h(\varsigma_i) = \begin{cases} \tilde{\mu}_{=} + \left(\varsigma_i - \frac{1}{nQ_{=}} \sum_{i \in I_{=}} \varsigma_i\right) \sqrt{\frac{nQ_{=}\tilde{\Sigma}_{=}}{\sum_{i \in I_{=}} \left(\varsigma_i - \frac{1}{nQ_{=}} \sum_{i' \in I_{=}} \varsigma_{i'}\right)^2}} & \text{if } i \in I_{=} \\ \tilde{\mu}_{\neq} + \left(\varsigma_i - \frac{1}{n(1-Q_{=})} \sum_{i \in I_{\neq}} \varsigma_i\right) \sqrt{\frac{n(1-Q_{=})\tilde{\Sigma}_{\neq}}{\sum_{i \in I_{\neq}} \left(\varsigma_i - \frac{1}{n(1-Q_{=})} \sum_{i' \in I_{\neq}} \varsigma_{i'}\right)^2}} & \text{if } i \in I_{\neq} \end{cases} \tag{14}$$

where $I_{=}$ and $I_{\neq}$ are the indices of correct and incorrect trials, respectively (and n is the total number of trials). **Eq. 14** ensures that the sample moments of the estimated normalized cumulative perturbations $\tilde{\eta}_i = h(\varsigma_i)$ match **Eqs 11, 12**, irrespective of the dummy variable ς. This also implies that the effective degrees of freedom of the constrained model are in fact lower than what the native self-consistency function would suggest.

b. Prior constraints on native DDM parameters.

In addition, one may want to introduce the following prior constraints on the native DDM parameters:

- The bound's height b is necessarily positive. This positivity constraint can be enforced by replacing b with a non-bounded parameter φ_1, which relates to the bound's height through the following mapping: $b = \exp(\varphi_1)$. We note that parameters ω of collapsing bounds $b_\omega(t)$ may not have to obey such positivity constraint.
- The standard deviation σ is necessarily positive. Again, this can be enforced by replacing it with the following mapped parameter φ_2: $\sigma = \exp(\varphi_2)$.
- The non-decision time T_{ND} is necessarily positive and smaller than the minimum observed reaction time. This can be enforced by replacing the native non-decision time with the following mapped parameter φ_3: $T_{ND} = \min(RT)s(\varphi_3)$, where $s(\cdot)$ is the standard sigmoid mapping.
- The initial bias x_0 is necessarily constrained between $-b$ and b. This can be enforced by replacing the native initial condition with the following mapped parameter φ_4: $x_0 = \exp(\varphi_1)[2s(\varphi_4) - 1]$.
- In principle, the drift rate v can be either positive or negative. However, its magnitude is necessarily smaller than $\frac{b+|x_0|}{\min(RT)-T_{ND}}$, which corresponds to its "ballistic" limit (see **Supplementary Appendix S6** for more details). This can be enforced by replacing the native drift rate with the following mapped parameter φ_5: $v = \frac{[1+|2s(\varphi_4)-1|]\exp(\varphi_1)}{\min(RT)[1-s(\varphi_3)]}[2s(\varphi_5) - 1]$.

Here again, we use the set of dummy variables $\varphi_{1:5}$ in lieu of native DDM parameters.

The statistical efficiency of the ensuing overcomplete approach can be evaluated by simulating RT and choice data under different settings of the DDM parameters, and then comparing estimated and simulated parameters. Below, we use such recovery analysis to compare the overcomplete approach with standard DDM fitting procedures.

c. Accounting for predictable trial-by-trial RT variability.

Critically, the above overcomplete approach can be extended to ask whether trial-by-trial variations in DDM parameters explain trial-by-trial variations in observed RT, above and beyond the impact of the random perturbation term in **Eq. 7**. For example, one may want to assess whether predictable variations in e.g., the drift term, accurately predict variations in RT data. This kind of questions underlies many recent empirical studies of human and/or animal decision making. In the context of perceptual decision making, the drift rate is assumed to derive from the strength of momentary evidence, which is controlled experimentally and varies in a trial-by-trial fashion (Huk and Shadlen, 2005; Bitzer et al., 2014). A straightforward extension of this logic to value-based decisions implies that the drift rate should vary in proportion to the value difference between alternative options (Krajbich et al., 2010; De Martino et al., 2012; Lopez-Persem et al., 2016). In both cases, a prediction for drift rate variations across trials is available, which is likely to induce trial-by-trial variations in choice and RT data. Let D be a known predictor variable, which is expected to capture trial-by-trial variations in some DDM parameter (e.g., the drift rate). One may then alter the self-consistency equation such that DDM parameters are treated as affine functions of trial-by-trial predictors (e.g., $v_i \triangleq v_0 + v_1 D_i$), and exploit trial-by-trial variations in response times to fit the ensuing offset and slope parameters (here, v_0 and v_1). Alternatively, one can simply set the drift rate to the predictor variable (i.e., assume *a priori* $v_0 = 0$ and $v_1 = 1$), which is currently the favorite approach in the field. As we will see below, this significantly improves model identifiability for the remaining parameters. This is because trial-by-trial variations in the drift rate will only accurately predict trial-by-trial variations in response time data if the remaining parameters are correctly set. This is just an example of course, and one can see how easily any prior dependency to a predictor variable could be accounted for.

The critical point here is that the overcomplete approach can exploit predictable trial-by-trial variations in RT data to improve the inference on model parameters.

PARAMETER RECOVERY ANALYSIS

In what follows, we use numerical simulations to evaluate the approach's ability to recover DDM parameters. Our parameter recovery analyses proceed as follows. First, we sample 1,000 sets of model parameters $\varphi_{1:5}$ under some arbitrary distribution. Second, for each of these parameter, we simulate a series of N = 200 DDM trials according to **Eq. 2** above. Third, we fit the DDM to each series of simulated reaction times (200 data points) and extract parameter estimates. Last, we compare simulated and estimated parameters to each other. In particular, we measure the relative estimation error for each DDM parameter. We also quantify potential non-identifiability issues using so-called recovery matrices and the ensuing identifiability index. We note that details regarding parameter recovery analyses can be found in **Supplementary Appendix S4** of this manuscript (along with definitions of the relative estimation error *REE*, recovery matrices and identifiability index ΔV).

To begin with, we will focus on "vanilla" DDMs, because they provide a fair benchmark for parameter estimation methods. In this context, we will compare the overcomplete approach with two established methods (Moens and Zenon, 2017; Boehm et al., 2018), namely: the "method of moments" and the "method of trial means". These methods are summarized in **Supplementary Appendixes S2, S3**, respectively. In brief, the former attempts to match empirical and theoretical moments of RT data. We expect this method to perform best when DDM parameters are fixed across trials. The latter rather attempts to match raw trial-by-trial RT data to trial-by-trial theoretical RT means. This will be most reliable when DDM parameters (e.g., the drift rate) vary over trials. Note that, in all cases, we inserted the prior constraints on DDM parameters given in *An Overcomplete Likelihood Approach to DDM Inversion* (section b) above, along with standard normal priors on unmapped parameters $\varphi_{1:5}$. We will therefore compare the ability of these methods to recover DDM parameters (i) when no parameter is fixed (full parameter set), (ii) when the drift rate is fixed, and (iii) when drift rates vary over trials.

Finally, we perform a parameter recovery analysis in the context of a generalized DDM, which includes collapsing bounds. This will serve to demonstrate the flexibility and robustness of the overcomplete approach.

a. Vanilla DDM: recovery analysis for the full parameter set.

First, we compare the three approaches when all DDM parameters have to be estimated. This essentially serves as a reference point for the other recovery analyses. The ensuing recovery analysis is summarized in **Figure 7** below, in terms of the comparison between simulated and estimated parameters.

Unsurprisingly, parameter estimates depend on the chosen estimation method, i.e. different methods exhibit distinct estimation errors structures. In addition, estimated and simulated parameters vary with similar magnitudes, and no systematic estimation bias is noticeable. It turns out that, in this setting, estimation error is minimal for the method of moments, which exhibits lower error than both the overcomplete approach (mean error difference: $\Delta\log(REE) = 0.27 \pm 0.03$, $p < 10^{-4}$, two-sided F-test) and the method of moments (mean error difference: $\Delta\log(REE) = 0.26 \pm 0.02$, $p < 10^{-4}$, two-sided F-test). However, the overcomplete approach and the method of trial means yield comparable estimation errors (mean error difference: $\Delta\log(REE) = 0.006 \pm 0.04$, $p = 0.88$, two-sided F-test).

Now, although estimation errors enable a coarse comparison of methods, it does not provide any quantitative insight regarding potential non-identifiability issues. We address this using recovery matrices (see **Supplementary Appendix S4**), which are shown on **Figure 8** below.

None of the estimation methods is capable of perfectly identifying DDM parameters (except T_{ND}), i.e., all methods exhibit strong non-identifiability issues. In particular, variations in the perturbations' standard deviation σ are partially confused with variations in the bound's height b, and reciprocally. This is because increasing both at the same time leaves RT trial-by-trial variability unchanged. Therefore, RT produced under strong neural perturbations can be equally well explained with a small bound height (and reciprocally). Interestingly, drift rate estimates are the least reliable: though their amount of "correct variability" is decent for the method of moments (45.3%), it is very low for both the overcomplete approach (5.3%) and the method of trial means (7.5%). If anything, non-identifiability issues are strongest for the overcomplete approach, which also exhibits weak "correct variability" for initial conditions (5.1%).

b. Vanilla DDM: recovery analysis with a fixed drift rate.

In fact, we expect non-identifiability issues of this sort, which were already highlighted in early DDM studies (Ratcliff, 1978). Note that this basic form of non-identifiability is easily disclosed from the self-consistency equation, which is invariant to a rescaling of all DDM parameters (except T_{ND}). In other terms, response times are left unchanged if all these parameters are rescaled by the same amount. Although this problematic invariance would disappear if a single DDM parameter was fixed rather than fitted, other non-identifiability issues may still hamper DDM parameter recovery. To test this, we re-performed the above parameter recovery analysis, but this time informing estimation methods about the drift rate, which was set to its simulated value. We note that such arbitrary reduction of the parameter space is routinely performed, as it was already suggested in seminal empirical applications of the DDM (Ratcliff, 1978). **Figure 9** below summarizes the ensuing comparison between simulated and estimated parameters.

Comparing **Figures 7**, **9** provides a clear insight regarding the impact of reducing the DDM's parameter space. In brief, estimation errors decrease for all methods, which seem to provide much more reliable parameter estimates. The method of moments still yields the most reliable parameter estimates,

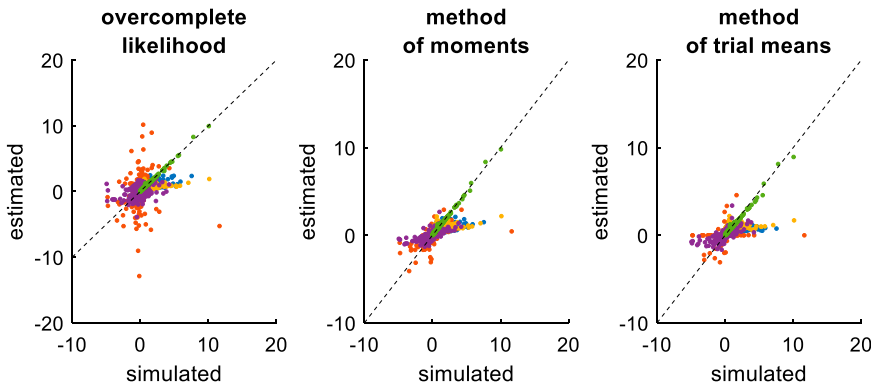


FIGURE 7 | Comparison of simulated and estimated DDM parameters (full parameter set). Left panel: Estimated parameters using the overcomplete approach (y-axis) are plotted against simulated parameters (x-axis). Each dot is a monte-carlo simulation and different colors indicate distinct parameters (blue: σ, red: v, yellow: b, purple: x_0, green: T_{ND}). The black dotted line indicate the identity line (perfect estimation). Middle panel: Method of moments, same format as left panel. Right panel: Method of trial means, same format as left panel.

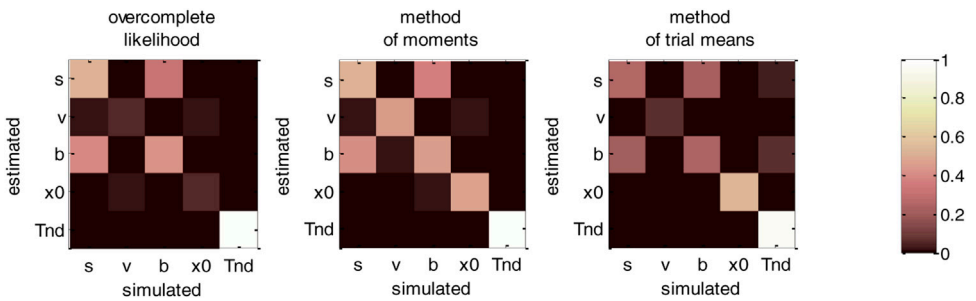


FIGURE 8 | DDM parameter recovery matrices (full parameter set). Left panel: overcomplete approach. Middle panel: method of moments. Right panel: Method of trial means. Each line shows the squared partial correlation coefficient between a given estimated parameter and each simulated parameter (across 1000 Monte-Carlo simulations). Note that perfect recovery would exhibit a diagonal structure, where variations in each estimated parameter is only due to variations in the corresponding simulated parameter. Diagonal elements of the recovery matrix measure "correct estimation variability", i.e., variations in the estimated parameters that are due to variations in the corresponding simulated parameter. In contrast, non-diagonal elements of the recovery matrix measure "incorrect estimation variability", i.e., variations in the estimated parameters that are due to variations in other parameters. Strong non-diagonal elements in recovery matrices thus signal pairwise non-identifiability issues.

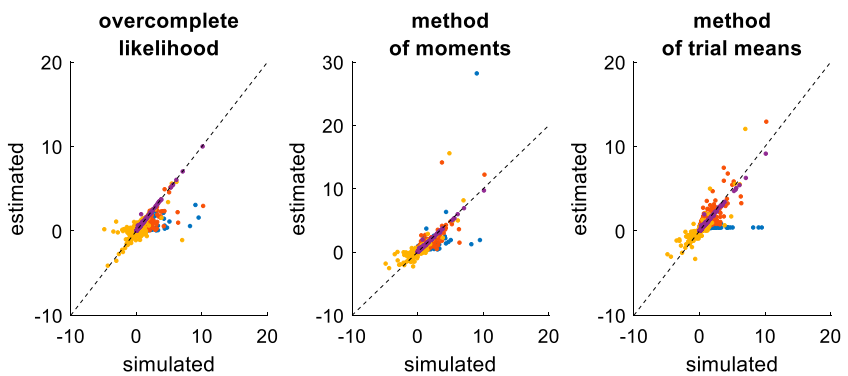


FIGURE 9 | Comparison of simulated and estimated DDM parameters (fixed drift rates). Same format as **Figure 7**, except for the color code in upper panels (blue: σ, red: b, yellow: x_0, purple: T_{ND}).

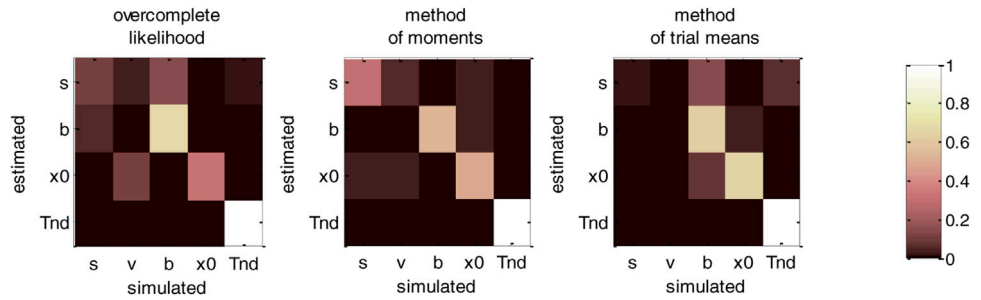


FIGURE 10 | DDM parameter recovery matrices (fixed drift rates). Same format as **Figure 8**, except that recovery matrices do not include the line that corresponds to the drift rate estimates. Note, however, that we still account for variations in the remaining estimated parameters that are attributable to variations in simulated drift rates.

eventually exhibiting lower error than the overcomplete approach (mean error difference: $\Delta\log(REE) = 0.21 \pm 0.03$, p = 0.04, two-sided F-test) and the method of trial means (mean error difference: $\Delta\log(REE) = 0.53 \pm 0.03$, $p < 10^{-4}$, two-sided F-test). In addition, the overcomplete approach yields lower estimation error than the method of trial means (mean error difference: $\Delta\log(REE) = 0.33 \pm 0.04$, $p < 10^{-4}$, two-sided F-test). The reason why the methods of trial means performs worst here is that it is blind to trial-by-trial variability in the data (beyond mean RT differences between the two decision outcomes). This is not the case however, for the two other methods.

We then evaluated non-identifiability issues using recovery matrices, which are summarized in **Figure 10** below.

Figure 10 clearly demonstrates an overall improvement in parameter identifiability (compare to **Figure 8**). In brief, most parameters are now identifiable, at least for the method of moments (which clearly performs best) and the overcomplete approach. Nevertheless, some weaker non-identifiability issues still remain, even when fixing the drift rate to its simulated value. For example, the overcomplete approach and the method of trial means still somehow confuse bound's heights with perturbations' standard deviations. More precisely, $\widehat{\sigma}$ shows unacceptably weak "correct variations" (overcomplete approach: 12.3%, method of trial means: 2.7%), when compared to "incorrect variations" due to the bound's height (overcomplete approach: 12.4%, method of trial means: 14.3%). Note that this does not hold for the method of moments, for which $\widehat{\sigma}$ shows strong "correct variations" (30.2%). Having said this, even the method of moments exhibit partial non-identifiability issues, in particular between perturbations' standard deviations and drift rates (incorrect variations: 4.1%).

We note that fixing another DDM parameter, e.g., the noise's standard deviation σ (instead of ν), would not change the relative merits of estimation methods in terms of parameter recovery. In other words, the above results are representative of the impact of fixing any DDM parameter. But situations where the drift rate is fixed can be directly compared with situations where one is attempting to exploit predictable drift rates trial-by-trial variations, which is the focus of the next section.

c. Vanilla DDM: recovery analysis with varying drift rates.

Now, accounting for predictable trial-by-trial variations in model parameters may, in principle, improve model identifiability. This is due to the fact that the net effect of each DDM parameter depends upon the setting of other parameters. Let us assume, for example, that the drift rate varies across trials according to some predictor variable (e.g., the relative evidence strength of alternative options in the context of perceptual decision making). The impact of other DDM parameters will not be the same, depending on whether the drift rate is high or low. In turn, there are fewer settings of these parameters that can predict trial-by-trial variations in RT data from variations in drift rate. To test this, we re-performed the recovery analysis, this time setting the drift rate according to a varying predictor variable, which is supposed to be known. The ensuing comparison between simulated and estimated parameters is summarized in **Figure 11** below.

On the one hand, the estimation error has now been strongly reduced, at least for the overcomplete approach and the method of trial means. On the other hand, estimation error has increased for the method of moments. This is because the method of moments confuses trial-by-trial variations that are caused by variations in drift rates with those that arise from the DDM's stochastic "neural" perturbation term. This is not the case for the overcomplete approach and the method of trial means. In turn, the method of moments now shows much higher estimation error than the overcomplete approach (mean error difference: $\Delta\log(REE) = 0.55 \pm 0.03$, $p < 10^{-4}$, two-sided F-test) or the method of trial means (mean error difference: $\Delta\log(REE) = 0.83 \pm 0.04$, $p < 10^{-4}$, two-sided F-test). Note that the latter eventually performs slightly better than the overcomplete approach (mean error difference: $\Delta\log(REE) = 0.28 \pm 0.03$, p = 0.04, two-sided F-test).

Figure 12 below then summarizes the evaluation of non-identifiability issues, in terms of recovery matrices.

For the overcomplete approach and the method of trial means, **Figure 12** shows a further improvement in parameter identifiability (compare to **Figures 8**, **10**). For these two methods, all parameters are now well identifiable ("correct variations" are always greater than 67.2% for all parameters), and no parameter estimate is strongly influenced by other simulated parameters. This is a simple example of the gain in

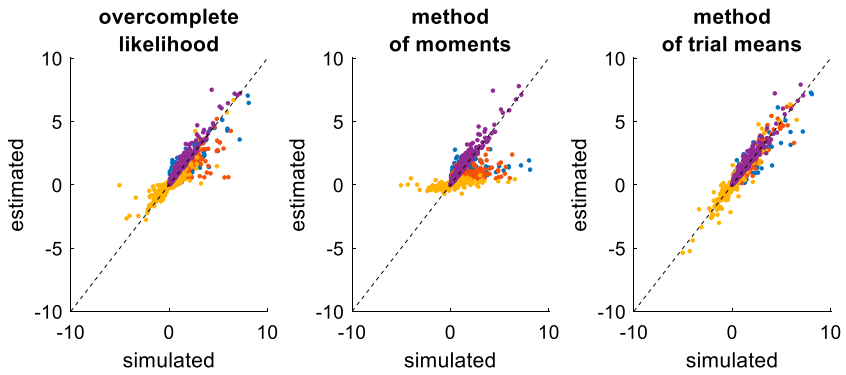


FIGURE 11 | Comparison of simulated and estimated DDM parameters (varying drift rates). Same format as **Figure 9**.

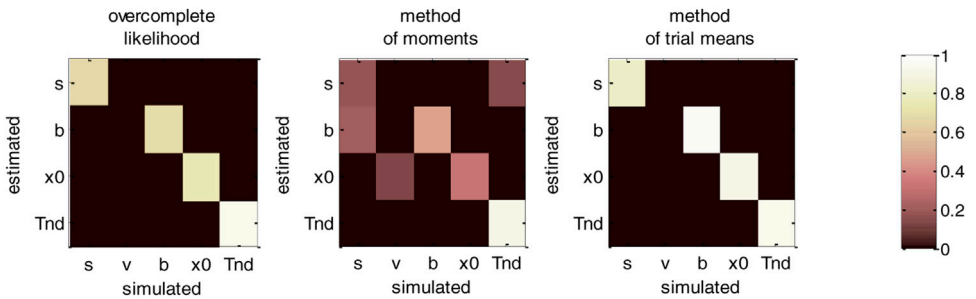


FIGURE 12 | DDM parameter recovery matrices (varying drift rates). Same format as **Figure 10**, except that fixed drift rates are replaced by their average across DDM trials.

statistical efficiency that result from exploiting known trial-by-trial variations in DDM model parameters. The situation is quite different for the method of moments, which exhibits clear non-identifiability issues for all parameters except the non-decision time. In particular, the bound's height is frequently confused with the perturbations' standard deviation (20.3% of "incorrect variations"), the estimate of which has become unreliable (only 17.6% of "correct variations").

We note that the gain in parameter recovery that obtains from exploiting predictable trial-by-trial variations in drift rates (with either the method of trial means or the overcomplete approach) does not generalize to situations where drift rates are defined in term of an affine transformation of some predictor variable (see *An Overcomplete Likelihood Approach to DDM Inversion* section. c above). This is because the ensuing offset and slope parameters would then need to be estimated along with other native DDM parameters. In turn, this would reintroduce identifiability issues similar or worse than when the full set of parameters have to be estimated (cf. *An Overcomplete Likelihood Approach to DDM Inversion* section.a). This is why people then typically fix another DDM parameter, e.g., the standard deviation σ (Ratcliff et al., 2016). But the risk of drawing erroneous conclusions, e.g., blindly interpreting differences due to σ in terms of differences in other DDM parameters, should invite modelers to be cautious with this kind of strategy.

d. Generalized DDM: recovery analysis with collapsing bounds.

We now consider generalized DDMs that include collapsing bounds. More precisely, we will consider a DDM where the bound $\widehat{b}_\omega(t)$ is exponentially decaying in time, i.e.: $\widehat{b}_\omega(t) = \exp(\omega_0 - \omega_1 t)$, where ω_0 and ω_1 control the bound's initial height and decay rate, respectively. This DDM variant reduces to the vanilla DDM when $\omega_1 \approx 0$, in which case the parameter ω_0 is formally identical to the vanilla bound's height b. When $\omega_1 \neq 0$ however, collapsing bounds induce a causal dependency between choice accuracy and response times that cannot be captured by the vanilla DDM (Zhang, 2012; Zhang et al., 2014; Hawkins et al., 2015; Tajima et al., 2016; Voskuilen et al., 2016).

In what follows, we report the results of a recovery analysis, in which data was simulated under the above generalized DDM (with drift rates varying across trials). We note that, under such generalized DDM variant, no analytical solution is available to derive RT moments. Applying the method of moments or the method of trial means to such generalized DDM variant thus involves either sampling schemes or numerical solvers for the underlying Fokker-Planck equation (Shinn et al., 2020). However, the computational cost of deriving trial-by-estimates of RT moments precludes routine data analysis using these methods,

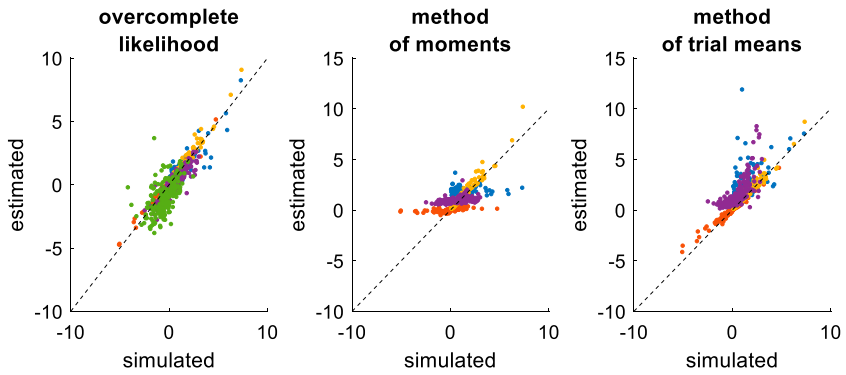


FIGURE 13 | Comparison of simulated and estimated DDM parameters (collapsing bounds). Same format as **Figure 9**, except that the left panel includes an additional parameter (w_1: green color), which controls the decay rate of DDM bounds.

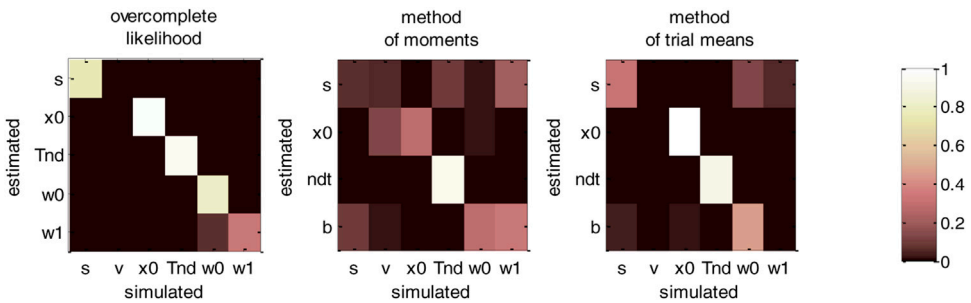


FIGURE 14 | DDM parameter recovery matrices (collapsing bounds). Same format as **Figure 12**, except that recovery matrices now also include the bound's decay rate parameter (w_1), in addition to the bound's initial height (w_0).

which is why most model-based studies are currently restricted to the vanilla DDM (Fengler et al., 2020). In turn, we do not consider here such computationally intensive extensions of the method of moments and/or method of trial means. In this setting, they thus do not rely on the correct generative model. The ensuing estimation errors and related potential identifiability issues should thus be interpreted in terms of the (lack of) robustness against simplifying modeling assumptions. This is not the case for the overcomplete approach, which bypasses this computational bottleneck and hence generalizes without computational harm to such DDM variants.

Figure 13 below summarizes the ensuing comparison between simulated and estimated parameters.

In brief, the overcomplete approach seems to perform as well as for non-collapsing bounds (see **Figure 11**). Expectedly however, the method of moments and the method of trial means do incur some reliability loss. Quantitatively, the overcomplete approach shows much smaller estimation error than the method of moments (mean error difference: $\Delta \log(REE) = 0.88 \pm 0.05$, $p < 10^{-4}$, two-sided F-test) or the method of trial means (mean error difference: $\Delta \log(REE) = 0.61 \pm 0.05$, $p < 10^{-4}$, two-sided F-test).

Figure 14 below then summarizes the ensuing evaluation of non-identifiability issues, in terms of recovery matrices.

For the overcomplete approach, **Figure 14** shows a similar parameter identifiability than **Figure 12**. In brief, all parameters of the generalized DDM are identifiable from each other (the amount of "correct variations" is 33.8% for the bound's decay parameter, and greater than 75.5% for all other parameters). This implies that including collapsing bounds does not impact parameter recovery with this method. This is not the case for the two other methods, however. In particular, the method of moments confuses the perturbations' standard deviation with the bound's decay rate (7.2% "correct variations" against 20.8% "incorrect variations"). This is also true, though to a lesser extent, for the method of trial means (31.6% "correct variations" against 5.4% "incorrect variations"). Again, these identifiability issues are expected, given that neither the method of moments nor the method of trial means (or, more properly, the variant that we use here) rely on the correct generative model. Maybe more surprising is the fact that these methods now exhibit non-identifiability issues w.r.t. parameters that they can, in principle, estimate. This exemplifies the sorts of interpretation issues that arise when relying on methods that neglect decision-relevant mechanisms. We will comment on this and related issues further in the Discussion section below.

e. Summary of recovery analyses.

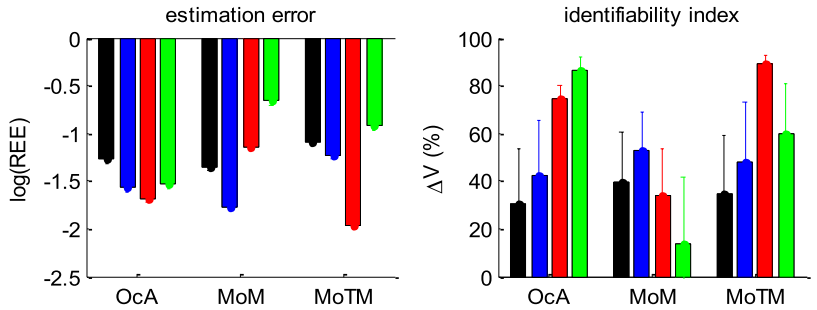


FIGURE 15 | Summary of DDM parameter recovery analyses. Left panel: The mean log relative estimation error *RRE* (*y*-axis) is shown for all methods (OcA: Overcomplete approach, MoM: Method of moments, MoTM: Method of trial means), and all simulation series (black: Full parameter set, blue: fixed drift rate, red: varying drift rates, green: Collapsing bounds). Right panel: The mean identifiability index ΔV (*y*-axis) is shown for all methods and all simulation series (same format as left panel). Note that the situation in which the full parameter set has to be estimated serves as a References point. To enable a fair comparison, both the estimation error and the identifiability index are computed for the parameter subset that is common to all simulation series (i.e.: The perturbations 'standard deviation σ, the bound's height b, the initial condition x_0, and the non-decision time T_{ND}).

Figure 15 below summarizes all our recovery analyses above, in terms of the average (log-) relative estimation error *REE* and the parameter identifiability index ΔV (cf. **Supplementary Appendix S4**).

Figure 15 enables a visual comparison of the impact of simulation series on parameter estimation methods. As expected, for the method of moments and the method of trial means, the most favorable situation (in terms of estimation error and identifiability) is when the drift rate is fixed and varying over trials, respectively. This is also when these methods perform best in relation to each other. All other situations are detrimental, and eventually yield estimation error and identifiability issues similar or worse than when the full parameter set has to be estimated. This is not the case for the overcomplete approach, which exhibits comparable estimation error and/or identifiability than the best method in all situations, except for collapsing bounds, where it strongly outperforms the two other methods. Here again, we note that parameter recovery for generalized DDMs may, in principle, be improved for the method of moments and/or the method of trial means. But extending these methods to generalized DDMs is beyond the scope of the current work.

APPLICATION TO A VALUE-BASED DECISION MAKING EXPERIMENT

To demonstrate the above overcomplete likelihood approach, we apply it to data acquired in the context of a value-based decision making experiment (Lopez-Persem et al., 2016). This experiment was designed to understand how option values are compared when making a choice. In particular, it tested whether agents may have prior preferences that create default policies and shape the neural comparison process.

Prior to the choice session, participants (n = 24) rated the likeability of 432 items belonging to three different domains (food, music, magazines). Each domain included four categories of 36 items. At that time, participants were unaware of these categories. During the choice session, subjects performed series of choices between two items, knowing that one choice in each domain would be randomly selected at the end of the experiment and that they would stay in the lab for another 15 min to enjoy their reward (listening to the selected music, eating the selected food and reading the selected magazine). Trials were blocked in a series of nine choices between items belonging to the same two categories within a same domain. The two categories were announced at the beginning of the block, such that subjects could form a prior or "default" preference (although they were not explicitly asked to do so). We quantified this prior preference as the difference between mean likeability ratings (across all items within each of the two categories). In what follows, we refer to the "default" option as the choice options that belonged to the favored category. Each choice can then be described in terms of choosing between the default and the alternative option.

Figure 16 below summarizes the main effects of a bias toward the default option (i.e., the option belonging to the favored category) in both choice and response time, above and beyond the effect of individual item values.

A simple random effect analysis based upon logistic regression shows that the probability of choosing the default option significantly increases with decision value, i.e. the difference V_{def}-V_{alt} between the default and alternative option values ($t = 8.4$, $dof = 23$, $p < 10^{-4}$). In addition, choice bias is significant at the group-level ($t = 8.7$, $dof = 23$, $p < 10^{-4}$). Similarly, RT significantly decreases with absolute decision value $|V_{def}$-$V_{alt}|$ ($t = 8.7$, $dof = 23$, $p < 10^{-4}$), and RT bias is significant at the group-level ($t = 7.4$, $dof = 23$, $p < 10^{-4}$).

To interpret these results, we fitted the DDM using the above overcomplete approach, when encoding the choice either (i) in terms of default versus alternative option (i.e., as is implicit on **Figure 10**) or (ii) in terms of right option versus left option. In what follows, we refer to the former choice frame as the "default/alternative" frame, and to the latter as the "native" frame. In both cases, the drift rate of each choice trial was set to the corresponding decision value (either V_{def}-V_{alt} or V_{right}-V_{left}). It turns out that within-subject estimates of σ, b and T_{ND} do not

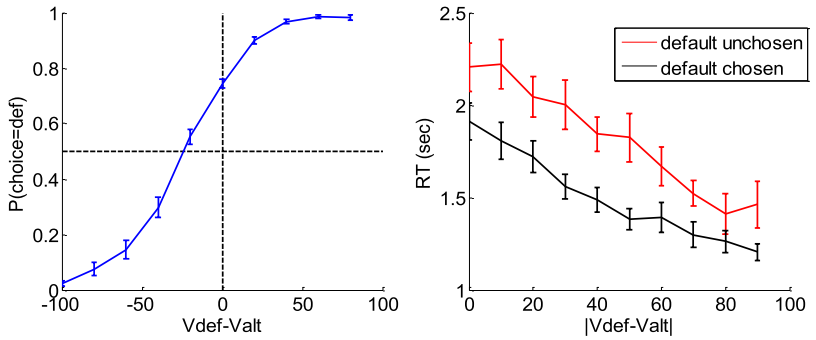


FIGURE 16 | Evidence for choice and RT biases in the default/alternative frame. Left: Probability of choosing the default option (*y*-axis) is plotted as a function of decision value V_{def}-V_{alt} (*x*-axis), divided into 10 bins. Values correspond to likeability ratings given by the subject prior to choice session. For each participant, the choice bias was defined as the difference between chance level (50%) and the observed probability of choosing the default option for a null decision value (i.e., when $V_{def} = V_{alt}$). Right: Response time RT (*y*-axis) is plotted as a function of the absolute decision value $|V_{def}$-$V_{alt}|$ (*x*-axis) divided into 10 bins, separately for trials in which the default option was chosen (black) or not (red). For each participant, the RT bias was defined as the difference between the RT intercepts (when $V_{def} = V_{alt}$) observed for each choice outcome.

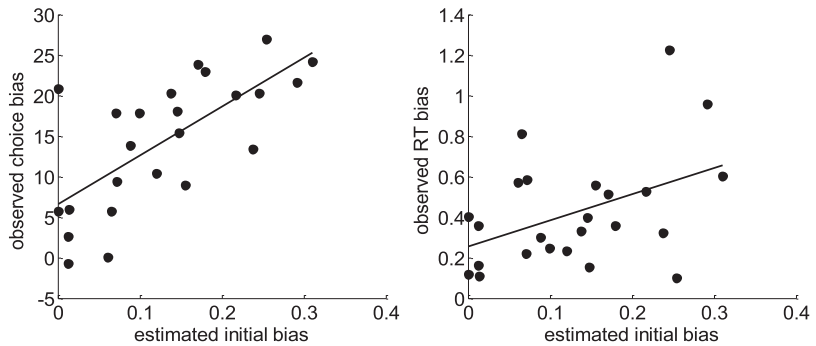


FIGURE 17 | Model-based analyses of choice and RT data. Left: For each participant, the observed choice bias (*y*-axis) is plotted as a function of the initial bias estimate $\hat{x}_0$ in the default/alternative frame (*x*-axis). Right: Same for the observed RT bias.

depend upon the choice frame. More precisely, the cross-subjects correlation of these estimates between the two choice frames is significant in all three cases (σ: r = 0.76, p < 10^{-4}; b: r = 0.82, p < 10^{-4}; T_{ND}: r = 0.94, p < 10^{-4}). This implies that inter-individual differences in σ, b and T_{ND} can be robustly identified, irrespective of the choice frame. However, the between-frame correlation is not significant for the initial bias x_0 (r = 0.29, p = 0.17). In addition, the initial bias is significant at the group level for the default/alternative frame (F = 45.2, dof = [1,23], p < 10^{-4}) but not for the native frame (F = 2.36, dof = [1,23], p = 0.14). In brief, the two choice frames only differ in terms of the underlying initial bias, which is only revealed in the default/alternative frame.

Now, we expect, from model simulations, that the presence of an initial bias induces both a choice bias, and a reduction of response times for default choices when compared to alternative choices (cf. upper-left and lower-right panels in **Figure 1**). The fact that $\hat{x}_0$ is significant in the default/alternative frame thus explains the observed choice and RT biases shown on **Figure 10**. But do inter-individual differences in $\hat{x}_0$ predict inter-individual differences in observed choice and RT biases? The corresponding statistical relationships are summarized on **Figure 17** below.

One can see that both pairs of variables are statistically related (choice bias: r = 0.70, p < 10^{-4}; RT bias: r = 0.44, p = 0.03). This is important, because this provides further evidence in favor of the hypothesis that people's covert decision frame facilitates the default option. Note that this could not be shown using the method of moments or the method of trial means, which were not able to capture these inter-individual differences (see **Supplementary Appendix S7** for details).

Finally, can we exploit model fits to provide a normative argument for why the brain favors a biased choice frame? Recall that, if properly set, the DDM can implement the optimal speed-accuracy tradeoff inherent in making online value-based decisions (Tajima et al., 2016). Here, it may seem that the presence of an initial bias would induce a gain in decision speed that would be overcompensated by the ensuing loss of accuracy. But in fact, the net tradeoff between decision speed and accuracy depends upon how the system sets the bound's height b. This is because b determines the demand for evidence before the system commits to a decision. More precisely, the system can favor decision accuracy by increasing b, or improve decision speed by decreasing b. We thus defined a measure $\widehat{e}$ of the

optimality of each participant's decisions, by comparing the speed-accuracy efficiency of her estimated DDM and the maximum speed-accuracy efficiency that can be achieved over alternative bound heights b (see **Supplementary Appendix SA5** below). This measure of optimality can be obtained either under the default-alternative frame or under the native frame. It turns out that the measured optimality of participants' decisions is significantly higher under the default/alternative frame than under the native frame ($\Delta\widehat{e} = 0.007 \pm 0.003$, t = 2.2, dof = 23, p = 0.02). In other words, participants' decisions appear more optimal under the default/alternative frame than under the native frame. We comment on possible interpretations of this result in the Discussion section below.

DISCUSSION

In this note, we have described an overcomplete approach to fitting the DDM to trial-by-trial RT data. This approach is based upon a self-consistency equation that response times obey under DDM models. It bypasses the computational bottleneck of existing DDM parameter estimation approaches, at the cost of augmenting the model with stochastic neural noise variables that perturb the underlying decision process. This makes it suitable for generalized variants of the DDM, which would not otherwise be considered for behavioral data analysis.

Strictly speaking, the DDM predicts the RT distribution conditional on choice outcomes. This is why variants of the method of moments are not optimal when empirical design parameters (e.g., evidence strength) are varied on a trial-by-trial basis. More precisely, one would need a few trial repetitions of empirical conditions (e.g., at least a few tens of trials per evidence strength) to estimate the underlying DDM parameters from the observed moments of associated RT distributions (Boehm et al., 2018; Ratcliff, 2008; Srivastava et al., 2016). Alternatively, one could rely on variants of the method of trial means to find the DDM parameters that best match expected and observed RTs (Fontanesi et al., 2019a; Fontanesi et al., 2019b; Gluth and Meiran, 2019; Moens and Zenon, 2017; Pedersen et al., 2017; Wabersich and Vandekerckhove, 2014). But this becomes computationally cumbersome when the number of trials is high and one wishes to use generalized variants of the DDM. This however, is not the case for the overcomplete approach. As with the method of trial means, its statistical power is maximal when design parameters are varied on a trial-by-trial basis. But the overcomplete approach does not suffer from the same computational bottleneck. This is because evaluating the underlying self-consistency equation (**Eqs. 7–9**) is much simpler than deriving moments of the conditional RT distributions (Broderick et al., 2009; Navarro and Fuss, 2009). In turn, the statistical added-value of the overcomplete approach is probably highest for analyzing data acquired with such designs, under generalized DDM variants.

We note that this feature of the overcomplete approach makes it particularly suited for learning experiments, where sequential decisions are based upon beliefs that are updated on a trial-by-trial basis from systematically varying pieces of evidence. In such contexts, existing modeling studies restrict the number of DDM parameters to deal with parameter recovery issues (Frank et al., 2015; Pedersen et al., 2017). This is problematic, since reducing the set of free DDM parameters can lead to systematic interpretation errors. In contrast, it would be trivial to extend the overcomplete approach to learning experiments without having to simplify the parameter space. We will pursue this in forthcoming publications.

Now what are the limitations of the overcomplete approach?

In brief, the overcomplete approach effectively reduces to adjusting DDM parameters such that RT become self-consistent. Interestingly, we derived the self-consistency equation without regard to the subtle dynamical degeneracies that (absorbing) bounds induce on stochastic processes (Broderick et al., 2009). It simply follows from noting that if a decision is triggered at time τ, then the underlying stochastic process has reached the bound (i.e., $x_\tau = \pm b$). This serves to identify the cumulative perturbation that eventually drove the system toward the bound. But a bound hit event at time τ is more informative about the history of the stochastic process than just its fate: it also tells us that the path did not cross the barrier before (i.e., $|x_t| < b \quad \forall t < \tau$). This disqualifies those sample paths whose first-passage time happens sooner, even though all barrier crossings are (by definition) "self-consistent". In retrospect, one may thus wonder whether the self-consistency equation may be suboptimal, in the sense of incurring some loss of information. Critically however, no information is lost about cumulative perturbations (or about DDM parameters). Although these are not sufficient to discriminate between the many sample paths that are compatible with a given RT, this is essentially irrelevant to the objective of the overcomplete approach. In turn, the existing limitations of the overcomplete approach lie elsewhere.

First and foremost, the self-consistency equation cannot be used to simulate data (recall that RTs appear on both the left- and right-hand sides of the equation). This restricts the utility of the approach to data analysis. Note however, that data simulations can still be performed using **Eq. 2**, once the model parameters have been identified. This enables all forms of posterior predictive checks and/or other types of model fit diagnostics (Palminteri et al., 2017). Second, the accuracy of the method depends upon the reliability of response time data. In particular, the recovery of the noise's standard deviation depends upon the accuracy of the empirical proxy for decision times (cf. second term in **Eq. 7**). In addition, the method inherits the potential limitations of its underlying parameter estimation technique: namely, the variational Laplace approach (Friston et al., 2007; Daunizeau, 2017). In particular, and as is the case for any numerical optimization scheme, it is not immune to multimodal likelihood landscapes. We note that this may result in non-identifiability issues of the sort that we have demonstrated here (cf., e.g., **Figures 8**, **10**). One cannot guarantee that this will not happen for some generalized DDM variant of interest. A possible diagnostic to this problem is to perform a systematic fit/sample/refit analysis to evaluate the stability of parameter estimates. In any case, we would advise to re-evaluate (and

report) parameter recovery for any novel DDM variant. Third, the computational cost of model inversion scales with the number of trials. This is because each trial has its own nuisance perturbation parameter. Note however, that the ensuing computational cost is many orders of magnitude lower than that of standard methods for generalized DDM variants. Fourth, proper bayesian model comparison may be more difficult. In particular, simulations show that a chance model always has a higher model evidence than the overcomplete model. This is another consequence of the overcompleteness of the likelihood function, which eventually pays a high complexity penalty cost in the context of Bayesian model comparison. Whether different DDM variants can be discriminated using the overcomplete approach is beyond the scope of the current work.

Let us now discuss the results of our model-based data analysis from the value-based decision making experiment (Lopez-Persem et al., 2016). Recall that we eventually provided evidence that peoples' decisions are more optimal under the default/alternative frame than under the native frame. Recall that this efficiency gain is inherited from the initial condition parameter x_0, which turns out be significant under the default/alternative frame. The implicit interpretation here is that the brain's decision system starts with a prior bias toward the default option. Critically however, we would have obtained the exact same results, would we have fixed the initial condition to zero but allowed upper and lower decision bounds to be asymmetrical. This is interesting, because it highlights a slightly different interpretation of our results. Under this alternative scenario, one would state that the brain's decision system is comparatively less demanding regarding the evidence that is required for committing to the default option. In turn, the benefit of lowering the bound for the default option may simply be to speed up decisions when evidence is congruent with default preferences, at the expense of slowing down incongruent decisions. Importantly, this strategy does not compromise decision accuracy if the incongruent decisions are rarer than the congruent ones (as is effectively the case in this experiment).

At this point, we would like to discuss potential neuroscientific applications of trial-by-trial estimates of "neural" perturbation terms. Recall that the self-consistency equation makes it possible to infer these neural noise variables from response times (cf. **Eq. 7** or 9). For the purpose of behavioral data analysis, where one is mostly interested in native DDM parameters, these are treated as nuisance variables. However, should one acquire neuroimaging data concurrently with behavioral data, one may want to exploit this unique feature of the overcomplete approach. In brief, estimates of "neural" perturbation terms moves the DDM one step closer to neural data. This is because DDM-based analysis of behavioral data now provides quantitative trial-by-trial predictions of an underlying neural variable. This becomes particularly interesting when internal variables (e.g., drift rates) are systematically varied over trials, hence de-correlating the neural predictor from response times. For example, in the context of fMRI investigations of value-based decisions, one may search for brain regions whose activity eventually perturbs the computation and/or comparison of options' values. This would extend the portfolio of recent empirical studies of neural noise perturbations to learning-relevant computations (Drugowitsch et al., 2016; Wyart and Koechlin, 2016; Findling et al., 2019). Reciprocally, using some variant of mediation analysis (MacKinnon et al., 2007; Lindquist, 2012; Brochard and Daunizeau, 2020), one may extract neuroimaging estimates of neural noise that can inform DDM-based behavioral data analysis. Alternatively, one may model neural and behavioral data in a joint and symmetrical manner, with the purpose of testing some predefined DDM variant (Rigoux and Daunizeau, 2015; Turner et al., 2015).

Finally, one may ask how generalizable the overcomplete approach is? Strictly speaking, one can evaluate the self-consistency equation under any DDM variant, as long as the mapping $z : x \rightarrow z(x)$ from the base random walk to the bound subspace is invertible (cf. **Eqs. 8, 9**). No such formal constraint exists for the dynamical form of the collapsing bound. This spans a family of DDM variants that is much broader than what is currently being used in the field (Fengler et al., 2020; Shinn et al., 2020). For example, this family includes decision models that trigger a decision when decision *confidence* reaches a bound (Tajima et al., 2016; Lee and Daunizeau, 2020). To the best of our knowledge, there is not a single example of existing DDM variants that does not belong to this class. Having said this, future extensions of the DDM framework may render the current overcomplete approach obsolete. Our guess is that such DDM improvements may then need to be informed with additional behavioral data, such as decision confidence (De Martino et al., 2012) and/or mental effort (Lee and Daunizeau, 2020), for which other kinds of self-consistency equations may be derived.

To conclude, we note that the code that is required to perform a DDM-based data analysis under the overcomplete approach will be made available soon from the VBA academic freeware https://mbb-team.github.io/VBA-toolbox/(Daunizeau et al., 2014).

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because they were not acquired by the authors. Requests to access the datasets should be directed to jean.daunizeau@inserm.fr.

ETHICS STATEMENT

Ethical review and approval was not required for the reuse of data from human participants in accordance with the local legislation. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

We would like to thank Alizée Lopez-Persem for providing us with the empirical data that serves to demonstrate our approach.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.531316/full#supplementary-material.

REFERENCES

Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J. A., Holmes, P., et al. (2011). Acquisition of decision making criteria: reward rate ultimately beats accuracy. *Atten. Percept. Psychophys.* 73, 640–657. doi:10.3758/s13414-010-0049-7

Beal, M. J. (2003). Variational algorithms for approximate Bayesian inference/. PhD Thesis. London: . University College London.

Bitzer, S., Park, H., Blankenburg, F., and Kiebel, S. J. (2014). Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. *Front. Hum. Neurosci.* 8, 102. doi:10.3389/fnhum.2014.00102

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., et al. (2018). Estimating across-trial variability parameters of the diffusion decision model: expert advice and recommendations. *J. Math. Psychol.* 87, 46–75. doi:10.1016/j.jmp.2018.09.004

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* 113, 700–765. doi:10.1037/0033-295x.113.4.700

Brochard, J., and Daunizeau, J. (2020). Blaming blunders on the brain: can indifferent choices be driven by range adaptation or synaptic plasticity? *BioRxiv*, 287714. doi:10.1101/2020.09.08.287714

Broderick, T., Wong-Lin, K. F., and Holmes, P. (2009). Closed-form approximations of first-passage distributions for a stochastic decision-making model. *Appl. Math. Res. Express* 2009, 123–141. doi:10.1093/amrx/abp008

Daunizeau, J., Adam, V., and Rigoux, L. (2014). VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioral data. *Plos Comput. Biol.* 10, e1003441. doi:10.1371/journal.pcbi.1003441

Daunizeau, J. (2017). The variational Laplace approach to approximate Bayesian inference. arXiv:1703.02089.

Daunizeau, J. (2019). Variational Bayesian modeling of mixed-effects. arXiv:1903.09003.

De Martino, B., Fleming, S. M., Garrett, N., and Dolan, R. J. (2012). Confidence in value-based choice. *Nat. Neurosci.* 16, 105–110. doi:10.1038/nn.3279

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *J. Neurosci.* 32, 3612–3628. doi:10.1523/jneurosci.4010-11.2012

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., and Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* 92, 1398–1411. doi:10.1016/j.neuron.2016.11.005

Fengler, A., Govindarajan, L. N., Chen, T., and Frank, M. J. (2020). *Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience*. BioRxiv. doi:10.1101/2020.11.20.392274

Findling, C., Chopin, N., and Koechlin, E. (2019). *Imprecise neural computations as source of human adaptive behavior in volatile environments*. BioRxiv, 799239.

Fontanesi, L., Gluth, S., Spektor, M. S., and Rieskamp, J. (2019a). A reinforcement learning diffusion decision model for value-based decisions. *Psychon. Bull. Rev.* 26, 1099–1121. doi:10.3758/s13423-018-1554-2

Fontanesi, L., Palminteri, S., and Lebreton, M. (2019b). Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling. *Cogn. Affect. Behav. Neurosci.* 19, 490–502. doi:10.3758/s13415-019-00723-1

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., et al. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *J. Neurosci.* 35, 485–494. doi:10.1523/jneurosci.2036-14.2015

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage* 34, 220–234. doi:10.1016/j.neuroimage.2006.08.035

Gluth, S., and Meiran, N. (2019). Leave-One-Trial-Out, LOTO, a general approach to link single-trial parameters of cognitive models to neural data. *eLife Sciences* 8. doi:10.7554/eLife.42607

Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.* 30, 535–574. doi:10.1146/annurev.neuro.29.051605.113038

Goldfarb, S., Leonard, N. E., Simen, P., Caicedo-Núñez, C. H., and Holmes, P. (2014). A comparative study of drift diffusion and linear ballistic accumulator models in a reward maximization perceptual choice task. *Front. Neurosci.* 8, 148. doi:10.3389/fnins.2014.00148

Grasman, R. P. P. P., Wagenmakers, E.-J., and van der Maas, H. L. J. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *J. Math. Psychol.* 53, 55–68. doi:10.1016/j.jmp.2009.01.006

Guevara Erra, R., Arbotto, M., and Schurger, A. (2019). An integration-to-bound model of decision-making that accounts for the spectral properties of neural data. *Sci. Rep.* 9, 8365. doi:10.1038/s41598-019-44197-0

Hanks, T., Kiani, R., and Shadlen, M. N. (2014). A neural mechanism of speed-accuracy tradeoff in macaque area LIP. *ELife* 3, e02260. doi:10.7554/elife.02260

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., and Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *J. Neurosci.* 35, 2476–2484. doi:10.1523/jneurosci.2410-14.2015

Huk, A. C., and Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *J. Neurosci.* 25, 10420–10436. doi:10.1523/jneurosci.4684-04.2005

Kloeden, P. E., and Platen, E. (1992). *Numerical solution of stochastic differential equations*. Berlin Heidelberg: Springer-Verlag.

Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* 13, 1292–1298. doi:10.1038/nn.2635

Lee, D., and Daunizeau, J. (2020). *Trading mental effort for confidence: the metacognitive control of value-based decision-making*. BioRxiv 837054.

Lindquist, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *J. Am. Stat. Assoc.* 107, 1297–1309. doi:10.1080/01621459.2012.695640

Lopez-Persem, A., Domenech, P., and Pessiglione, M. (2016). How prior preferences determine decision-making frames and biases in the human brain. *ELife* 5, e20317. doi:10.7554/elife.20317

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation analysis. *Annu. Rev. Psychol.* 58, 593. doi:10.1146/annurev.psych.58.110405.085542

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., and Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgm. Decis. Mak.* 5, 437–449.

Moens, V., and Zenon, A. (2017). Variational treatment of trial-by-trial drift-diffusion models of behavior. *BioRxiv 220517*.

Navarro, D. J., and Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *J. Math. Psychol.* 53, 222–230. doi:10.1016/j.jmp.2009.02.003

Newey, W. K., and West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *Int. Econ. Rev.* 28, 777–787. doi:10.2307/2526578

Osth, A. F., Bora, B., Dennis, S., and Heathcote, A. (2017). Diffusion vs. linear ballistic accumulation: different models, different conclusions about the slope of the zROC in recognition memory. *J. Mem. Lang.* 96, 36–61. doi:10.1016/j.jml.2017.04.003

Palminteri, S., Wyart, V., and Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* 21, 425–433. doi:10.1016/j.tics.2017.03.011

Pedersen, M. L., Frank, M. J., and Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychon. Bull. Rev.* 24, 1234–1251. doi:10.3758/s13423-016-1199-y

Pedersen, M. L., and Frank, M. J. (2020). Simultaneous hierarchical bayesian parameter estimation for reinforcement learning and drift diffusion models: a tutorial and links to neural data. *Comput. Brain Behav.* 3, 458–471. doi:10.1007/s42113-020-00084-w

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev.* 85, 59–108. doi:10.1037/0033-295x.85.2.59

Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput.* 20, 873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends Cogn. Sci.* 20, 260–281. doi:10.1016/j.tics.2016.01.007

Ratcliff, R. (2008). The EZ diffusion method: too EZ? *Psychon. Bull. Rev.* 15, 1218–1228. doi:10.3758/pbr.15.6.1218

Ratcliff, R., and Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. *Psychon. Bull. Rev.* 9, 438–481. doi:10.3758/bf03196302

Resulaj, A., Kiani, R., Wolpert, D. M., and Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature* 461, 263–266. doi:10.1038/nature08275

Rigoux, L., and Daunizeau, J. (2015). Dynamic causal modeling of brain-behavior relationships. *NeuroImage* 117, 202–221. doi:10.1016/j.neuroimage.2015.05.041

Shinn, M., Lam, N. H., and Murray, J. D. (2020). A flexible framework for simulating and fitting generalized drift-diffusion models. *ELife* 9, e56938. doi:10.7554/elife.56938

Srivastava, V., Holmes, P., and Simen, P. (2016). Explicit moments of decision times for single- and double-threshold drift-diffusion processes. *J. Math. Psychol.* 75, 96–109. doi:10.1016/j.jmp.2016.03.005

Tajima, S., Drugowitsch, J., and Pouget, A. (2016). Optimal policy for value-based decision-making. *Nat. Commun.* 7, 12400. doi:10.1038/ncomms12400

Turner, B. M., van Maanen, L., and Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: the neural drift diffusion model. *Psychol. Rev.* 122, 312–336. doi:10.1037/a0038894

Vandekerckhove, J., and Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: a DMAT primer. *Behav. Res.* 40, 61–72. doi:10.3758/brm.40.1.61

Voskuilen, C., Ratcliff, R., and Smith, P. L. (2016). Comparing fixed and collapsing boundary versions of the diffusion model. *J. Math. Psychol.* 73, 59–79. doi:10.1016/j.jmp.2016.04.008

Voss, A., and Voss, J. (2007). Fast-dm: a free program for efficient diffusion model analysis. *Behav. Res. Methods* 39, 767–775. doi:10.3758/bf03192967

Wabersich, D., and Vandekerckhove, J. (2014). Extending JAGS: a tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behav. Res.* 46, 15–28. doi:10.3758/s13428-013-0369-3

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., and Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychon. Bull. Rev.* 15, 1229–1235. doi:10.3758/pbr.15.6.1229

Wagenmakers, E.-J., van der Maas, H. L. J., and Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychon. Bull. Rev.* 14, 3–22. doi:10.3758/bf03194023

Wiecki, T. V., Sofer, I., and Frank, M. J. (2013). HDDM: hierarchical bayesian estimation of the drift-diffusion model in Python. *Front. Neuroinformatics* 7, 14. doi:10.3389/fninf.2013.00014

Wyart, V., and Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. *Curr. Opin. Behav. Sci.* 11, 109–115. doi:10.1016/j.cobeha.2016.07.003

Zhang, J. (2012). The effects of evidence bounds on decision-making: theoretical and empirical developments. *Front. Psychol.* 3, 263. doi:10.3389/fpsyg.2012.00263

Zhang, S., Lee, M. D., Vandekerckhove, J., Maris, G., and Wagenmakers, E.-J. (2014). Time-varying boundaries for diffusion models of decision making and response time. *Front. Psychol.* 5, 1364. doi:10.3389/fpsyg.2014.01364

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

REVIEW
published: 21 May 2021
doi: 10.3389/frai.2021.530937

Neuronal Sequence Models for Bayesian Online Inference

Sascha Frölich, Dimitrije Marković and Stefan J. Kiebel*

Department of Psychology, Technische Universität Dresden, Dresden, Germany

Various imaging and electrophysiological studies in a number of different species and brain regions have revealed that neuronal dynamics associated with diverse behavioral patterns and cognitive tasks take on a sequence-like structure, even when encoding stationary concepts. These neuronal sequences are characterized by robust and reproducible spatiotemporal activation patterns. This suggests that the role of neuronal sequences may be much more fundamental for brain function than is commonly believed. Furthermore, the idea that the brain is not simply a passive observer but an active predictor of its sensory input, is supported by an enormous amount of evidence in fields as diverse as human ethology and physiology, besides neuroscience. Hence, a central aspect of this review is to illustrate how neuronal sequences can be understood as critical for probabilistic predictive information processing, and what dynamical principles can be used as generators of neuronal sequences. Moreover, since different lines of evidence from neuroscience and computational modeling suggest that the brain is organized in a functional hierarchy of time scales, we will also review how models based on sequence-generating principles can be embedded in such a hierarchy, to form a generative model for recognition and prediction of sensory input. We shortly introduce the Bayesian brain hypothesis as a prominent mathematical description of how online, i.e., fast, recognition, and predictions may be computed by the brain. Finally, we briefly discuss some recent advances in machine learning, where spatiotemporally structured methods (akin to neuronal sequences) and hierarchical networks have independently been developed for a wide range of tasks. We conclude that the investigation of specific dynamical and structural principles of sequential brain activity not only helps us understand how the brain processes information and generates predictions, but also informs us about neuroscientific principles potentially useful for designing more efficient artificial neuronal networks for machine learning tasks.

Keywords: neuronal sequences, Bayesian inference, generative models, Bayesian brain hypothesis, predictive coding, hierarchy of time scales, recurrent neural networks, spatiotemporal trajectories

OPEN ACCESS

Edited by:
Bertram Müller-Myhsok, Max Planck Institute of Psychiatry (MPI), Germany

Reviewed by:
Hazem Toutounji, University of Nottingham, United Kingdom
Philipp Georg Sämann, Max Planck Institute of Psychiatry, Germany

****Correspondence:***
Sascha Frölich
sascha.froelich@tu-dresden.de

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *30 January 2021*
Accepted: *13 April 2021*
Published: *21 May 2021*

Citation:
Frölich S, Marković D and Kiebel SJ (2021) Neuronal Sequence Models for Bayesian Online Inference. Front. Artif. Intell. 4:530937. doi: 10.3389/frai.2021.530937

1. INTRODUCTION

In the neurosciences, one important experimental and theoretical finding of recent years was that many brain functions can be described as predictive (Rao and Ballard, 1999; Pastalkova et al., 2008; Friston and Kiebel, 2009; Aitchison and Lengyel, 2017). This means that the brain not only represents current states of the environment but also potential states of the future to adaptively select its actions and behavior. For such predictions, one

important feature of neuronal dynamics is their often-observed sequence-like structure. In this review, we will present evidence that sequence-like structure in neuronal dynamics is found over a wide range of different experiments and different species. In addition, we will also review models for such sequence-like neuronal dynamics, which can be used as generative models for Bayesian inference to compute predictions. To familiarize readers of different backgrounds with each of these topics, we first briefly give an overview of the topics of sequences, predictions, hierarchical structure, the so-called Bayesian brain hypothesis and provide a more precise definition of the kind of sequence-like neuronal dynamics that we consider in this review.

1.1. Sequences in the Brain

The brain is constantly receiving spatiotemporally structured sensory input. This is most evident in the auditory domain where, when listening to human speech, the brain receives highly structured, sequential input in the form of phonemes, words, and sentences (Giraud and Poeppel, 2012). Furthermore, even in situations which apparently provide only static sensory input, the brain relies on spatiotemporally structured coding. For example, when observing a static visual scene, the eyes constantly perform high-frequency micro-oscillations and exploratory saccades (Martinez-Conde et al., 2004; Martinez-Conde, 2006), which renders the visual input spatiotemporally structured, and yet the visual percepts appear stationary. Another example is olfaction, where in animal experiments, it has been shown that neurons in the olfactory system respond to a stationary odor with an elaborate temporal coding scheme (Bazhenov et al., 2001; Jones et al., 2007). In the state space of those neurons, their activity followed a robust and reproducible trajectory, a neuronal sequence (see **Table 1**), which was specific to the presented odor. Similarly, in a behavioral experiment with monkeys, spatial information of an object was encoded by a dynamical neural code, although the encoded relative location of the object remained unchanged (Crowe et al., 2010). In other words, there is evidence that the brain recognizes both dynamic and static entities in our environment on the basis of sequence-like encoding.

Neuronal sequences have been reported in a wide range of experimental contexts. For example, in the hippocampus of mice and rats (MacDonald et al., 2011; Pastalkova et al., 2008; Bhalla, 2019; Skaggs and McNaughton, 1996; Dragoi and Tonegawa, 2011), the visual cortex of cats and rats (Kenet et al., 2003; Ji and Wilson, 2007), the somatosensory cortex of mice (Laboy-Juárez et al., 2019), the parietal cortex of monkeys and mice (Crowe et al., 2010; Harvey et al., 2012), the frontal cortex of monkeys (Seidemann et al., 1996; Abeles et al., 1995; Baeg et al., 2003), the gustatory cortex of rats (Jones et al., 2007), the locust antennal lobe (Bazhenov et al., 2001), specific song-related areas in the brain of songbirds (Hahnloser et al., 2002), and the amygdala of monkeys (Reitich-Stolero and Paz, 2019), among others. Even at the cellular level, there is evidence of sequence-processing capacities of single neurons (Branco et al., 2010). Neuronal sequences seem to serve a variety of different purposes. While sequences in specific brain regions drive the spatiotemporal motor patterns during behavior like birdsong rendition (Hahnloser et al., 2002) (**Figure 1B**), in other studies of different brain areas and different species, neuronal sequences were found to encode stationary stimuli (Seidemann et al., 1996; Bazhenov et al., 2001) and spatial information (Crowe et al., 2010), to represent past experience (Skaggs and McNaughton, 1996) (see also **Figure 1A**), and to be involved with both working memory and memory consolidation (MacDonald et al., 2011; Harvey et al., 2012; Skaggs and McNaughton, 1996). Behaviorally relevant neuronal sequences were reported to occur before the first execution of a task (Dragoi and Tonegawa, 2011), and in some behavioral tasks sequences were found to be predictive of future behavior (Abeles et al., 1995; Pastalkova et al., 2008).

As these findings show, neuronal sequences can be measured in different species, in different brain areas and at different levels of observation, where the expression of these sequences depends on the measurement and analysis method. A neuronal sequence can appear as the successive spiking of neurons (**Figures 1A,B**), or the succession of more abstract compound states (**Figure 1C**), or in yet different forms, depending on the experimental approach. For example, evidence for sequences can also be found with non-invasive cognitive neuroscience methods like magnetoencephalography (MEG) as shown in **Figure 1D**. Given

TABLE 1 | Glossary.

Neuronal sequence	Spatiotemporal patterns of neuronal activity that encode stimulus properties, abstract concepts, or motion signals (see **Figure 1**). Can be described by a specific, sequential trajectory in the so-called state space of the system, see also **Figure 3** for an example.
State space/Phase space	A multidimensional space that encompasses all possible states a system can be in. Every possible state is defined by a unique point in the space.
Continuodiscrete dynamics/Trajectory	Reproducible spatiotemporal trajectories characterized by discrete points in state space (see **Figure 3**).
Winnerless Competition (WLC)	Type of dynamic behavior of a system where the system shortly settles into a stable or metastable state before being forced away from it (by internal or external mechanisms) (see **Figures 3**, **6**).
Metastable state/Saddle state	A state in the state space of a dynamical system. A metastable state of a system is stable in some directions and unstable in others. A saddle point is a metastable point where the first derivative vanishes.
Stable heteroclinic channel (SHC)	Type of dynamic behavior of a system where the system goes through a succession of saddle points (metastable states) forming heteroclinic state-space trajectories (orbits). Importantly, small deviations from those trajectories will not diverge away from the heteroclinic orbit. See section 2.2.2.
Heteroclinic orbit/Trajectory	A path in the state space of a system that connects two equilibrium points.
Limit cycle	Attractor type occurring in some complex dynamical systems. Closed, continuous trajectory in state space with fixed period and amplitude. The regular firing behavior of neurons can be described by limit cycle behavior. See section 2.2.1.
Synfire chain	A feed-forward neuronal network architecture. See section 2.1.

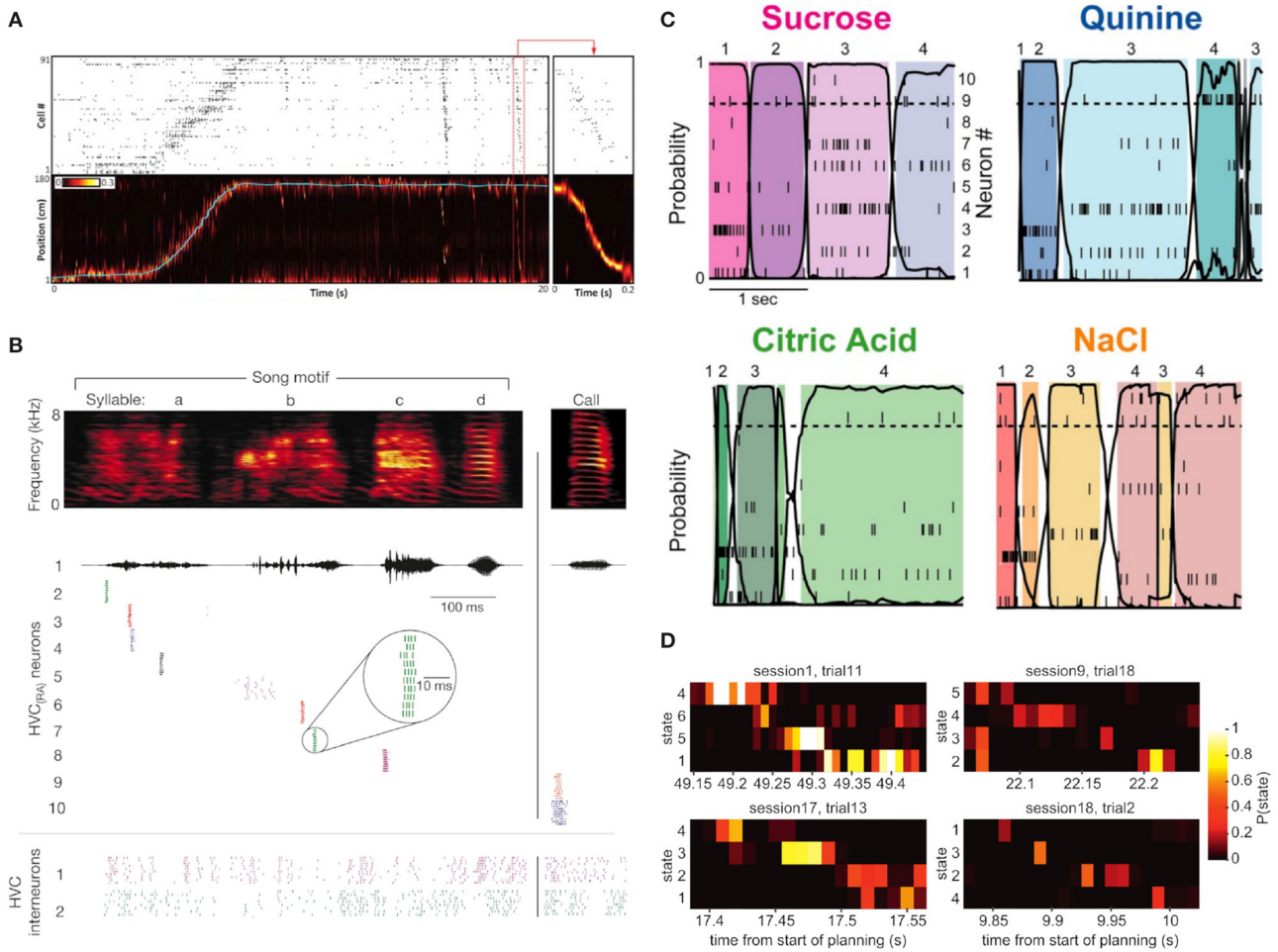


FIGURE 1 | Four illustrative examples of sequential neuronal activity in different paradigms and experimental contexts. **(A)** Sequential activation of rat hippocampal cells are found during action and in rest phases after the behavioral tasks. The top plot shows the spiking histogram of 91 hippocampal cells during a rat's trip along a physical track. The bottom panel shows the rat's actual position on the track (blue line) against the position inferred from the spiking pattern of its hippocampal cells. After the traversal of the track, hippocampal cells "replayed" their activation sequence in reverse during a short ripple event (red box, enlarged in the box on the right). Figure adapted from Pfeiffer (2020) (Copyright 1999–2019 John Wiley & Sons, Inc.). **(B)** Zebra finches are songbirds whose songs consist of highly consistent so-called song motifs. Here, the activations of ten different $HVC_{(RA)}$ neurons and two HVC interneurons in the HVC nucleus of the zebra finch brain during ten renditions of the same song motif are shown. $HVC_{(RA)}$ project from the HVC nucleus to the RA nucleus in the birdbrain, and exhibit precise and reproducible firing sequences during the rendition of a song. Adapted from Hahnloser et al. Hahnloser et al. (2002) with permission from Springer Nature. **(C)** Firing patterns of neurons in the gustatory cortex of rats *in vivo* when presented with four different odors. The sequential switching of states of a hidden Markov model (HMM, see section 3.1) was characteristic of the presented aroma. For each of the four odors, the different color hues represent different HMM states that were inferred based on the data. Adapted from Jones et al. (2007) (Copyright 2007 National Academy of Sciences, U.S.A.). **(D)** Evidence for fast sequence representation in human participants during planning of a trajectory through task state space, see Kurth-Nelson et al. (2016) for details. The four examples, each for a different participant, show evidence of brain activity, as measured with magnetoencephalography (MEG), to quickly transition through task state space with roughly 40 ms duration for each sequence element. Figure taken from Kurth-Nelson et al. (2016).

these very different appearances of experimentally observed neuronal sequences, it is clear that an answer to the question of "What is a neuronal sequence?" depends on the experimental setup. In the context of this article, we understand a "neuronal sequence" quite broadly as any kind of robust and reproducible spatiotemporal trajectory, where stimulus properties, abstract concepts, or motion signals are described by a specific trajectory in the state space of the system (see **Table 1**). The brain may use such trajectory representations, whose experimental expressions are measured as neuronal sequences, to form a basis for encoding the spatiotemporal structure of sensory stimuli (Buonomano and Maass, 2009) and the statistical dependencies between past, present, and future (Friston and Buzsáki, 2016). Here, we will review evidence for this type of encoding and discuss some of the implications for our understanding of the brain's capacity to perform probabilistic inference, i.e., recognition based on spatiotemporally structured sensory input.

1.2. Hierarchies in the Brain

The brain's structure and function are often described with reference to a hierarchical organization, which we will cover in more detail in section 3.2. Human behavior can be described as a hierarchically structured process (Lashley and Jeffress, 1951; Rosenbaum et al., 2007; Dezfouli et al., 2014), as can memory, where the grouping of information-carrying elements into chunks constitutes a hierarchical scheme (Bousfield, 1953; Miller, 1956; Fonollosa et al., 2015). Similarly, the perception and recognition of spatiotemporally structured input can be regarded as a hierarchical process. For example, percepts, such as the observation of a walking person can be regarded as percepts of higher order ("walking person"), as they emerge from the combination of simpler, lower order percepts, e.g., a specific sequence of limb movements. Critically, the concept "someone walking" is represented at a slower time scale as compared to the faster movements of individual limbs that constitute the walking. There is emerging evidence that the brain is structured and organized hierarchically along the relevant time scales of neuronal sequences (e.g., Murray et al., 2014; Hasson et al., 2008; Cocchi et al., 2016; Mattar et al., 2016; Gauthier et al., 2012; Kiebel et al., 2008). Such a hierarchy allows the brain to model the causal structure of its sensory input and form predictions at slower time scales ("someone walking") by representing trajectories capturing the dynamics of its expected spatiotemporal sensory input at different time scales, and by representing causal dependencies between time scales. This allows for inference about the causes of sensory input in the environment, as well as for inference of the brain's own control signals (e.g., motor actions). In this paper, we will review some of the experimental evidence and potential computational models for sequence generation and inference.

In the following section 1.3 we will first give a short introduction to the Bayesian brain hypothesis and the basic concept of the brain as a predictor of its environment. In section 1.4 we will go into more detail about the question "What is a sequence?" and will further discuss the trajectory representation. In section 2, we will provide an overview of several dynamical principles that might underlie the generation of neuronal trajectories in biological networks. Importantly, we are going to focus on general dynamical network principles that may underlie sequence generation, and which may differentiate types of sequence-generating networks. We are therefore not going to cover the vast field of sequence learning (e.g., Sussillo and Abbott, 2009; Tully et al., 2016; Lipton et al., 2015; Wörgötter and Porr, 2005), which mainly investigates neurobiologically plausible learning rules and algorithms that can lead to neuronal sequences, and thus possibly to the network types discussed in this article. In section 3, we review some approaches in which sequences are used to model recognition of sensory input. To highlight the relevance of sequence generators to a large variety of problems, we will visit methods and advances in computer science and machine learning, where structured artificial recurrent neural networks (RNNs) that are able to generate spatiotemporal activity patterns are used to perform a range of different computational tasks. This will however only serve as a rough and incomplete overview over some common machine learning methods, and we will not cover methods like Markov Decision Processes (Feinberg and Shwartz, 2012) and related approaches, as an overview of research on sequential decision making is beyond the scope of this review. Finally, we will briefly discuss functional hierarchies in the brain and in machine learning applications. A glossary of technical terms that we will use in the review can be found in **Table 1**.

1.3. The Bayesian Brain Hypothesis

Dating back to Hermann von Helmholtz in the 19th century, the idea that the brain performs statistical inference on its sensory input to infer the underlying probable causes of that same input (Helmholtz, 1867), started gaining considerable traction toward the end of the 20th century and had a strong influence on both computer science and neuroscience (Hinton and Sejnowski, 1983; Dayan et al., 1995; Wolpert et al., 1995; Friston, 2005; Friston et al., 2006; Beck et al., 2008; see also Rao and Ballard, 1999; Ernst and Banks, 2002; Körding and Wolpert, 2004). In particular, research into this interpretation of brain function led to the formulation of the Bayesian brain hypothesis (Knill and Pouget, 2004; Doya et al., 2007; Friston, 2010). The Bayesian brain hypothesis posits that aspects of brain function can be described as equivalent to Bayesian inference based on a causal generative model of the world, which models the statistical and causal regularities of the environment. In this framework, recognition is modeled as Bayesian inversion of the generative model, which assigns probabilities, that is, beliefs to different states of the world based on perceived sensory information. This process of Bayesian inference is hypothesized to be an appropriate basis for the mathematical description of most, if not all, brain functions (Friston, 2010; Knill and Pouget, 2004). Although the hypothesis that the brain is governed by Bayesian principles has met with criticism since human behavior does not always appear to be Bayes-optimal (Rahnev and Denison, 2018; Soltani et al., 2016), and because the definition of Bayes-optimality can be ambiguous (Colombo and Seriès, 2012), there is growing evidence that human behavior can indeed be explained by Bayesian principles (**Figure 2**) (Ernst and Banks, 2002; Körding and Wolpert, 2004; Weiss et al., 2002; Feldman, 2001), and that even phenomena like mental disorders might be explained by Bayesian mechanisms (Adams et al., 2013; Leptourgos et al., 2017; Fletcher and Frith, 2009) (see Knill and Pouget, 2004 and Clark, 2013 for reviews on the Bayesian brain hypothesis). How Bayesian inference is achieved in the human brain is an ongoing debate, and it has been proposed that the corresponding probabilities are encoded on a population level (Zemel et al., 1998; Beck et al., 2008) or on single-neuron level (Deneve, 2008).

Under the Bayesian view, model inversion, i.e., recognition, satisfies Bayes' theorem, which states that the optimal posterior belief about a state is proportional to the generative model's prior expectation about the state multiplied by the probability of the sensory evidence under the generative model. In Bayesian inference, prior expectation, posterior belief, and sensory evidence are represented as probability distributions and accordingly called *prior distribution*, *posterior distribution*, and *likelihood* (**Figure 2**). The posterior can be regarded as an updated version of the prior distribution, and will act as the prior

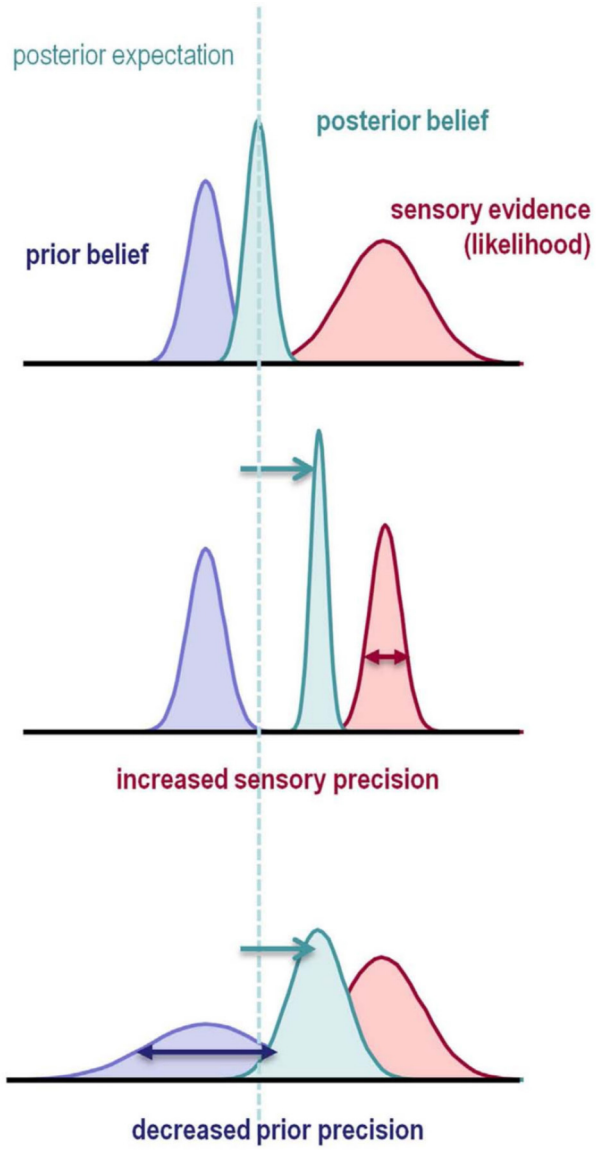


FIGURE 2 | Illustration of Bayesian Inference. The prior belief (blue) about a state is updated by sensory evidence (red) represented by the likelihood function. The updated belief is the posterior belief (turquoise), which will serve as the prior belief in the next updating step. Each row illustrates how the shape of the prior distribution and the likelihood influence the inference process. Both an increase in likelihood precision (inverse variance), and a decrease in prior precision result in a posterior belief which is more biased toward the sensory evidence. This is illustrated by a deviation of the posterior toward the sensory evidence and away from the prior belief (dashed line and arrows). In the Bayesian predictive coding framework (Friston and Kiebel, 2009; Rao and Ballard, 1999), inference naturally minimizes the prediction error, defined as the difference between expected and observed outcomes. Figure reprinted from Adams et al. (2013).

in the next inference step. Importantly, the prior is part of the generative model as different priors could lead to qualitatively different expectations (Gelman et al., 2017).

The quality of the inference, that is, the quality of the belief about the hidden states of the world, is dependent on the quality of the agent's generative model, and the appropriateness of a tractable (approximate) inference scheme. In this review paper, we suggest that good generative models of our typical environment should generate, that is, expect sequences, and that such a sequence-like representation of environmental dynamics is used to robustly perform tractable inference on spatiotemporally structured sensory data.

The theory of predictive coding suggests that the equivalent of an inversion of the generative model in the cortex is achieved in a hierarchical manner by error-detecting neurons which encode the difference between top-down predictions and sensory input (Friston and Kiebel, 2009; Rao and Ballard, 1999; Aitchison and Lengyel, 2017) (**Figure 2**). The fact that sequences in specific contexts appear to have predictive properties (Abeles et al., 1995; Pastalkova et al., 2008) is interesting in light of possible combinations of the frameworks of predictive coding and the Bayesian brain hypothesis (Knill and Pouget, 2004; Doya et al., 2007; Friston, 2010). One intriguing idea is that the brain's internal representations and predictions rely on sequences of neuronal activity (FitzGerald et al., 2017; Kiebel et al., 2009; Hawkins et al., 2009). Importantly, empirical evidence suggests that these approximate representations are structured in temporal and functional hierarchies (see sections 1.2 and 3.2) (Koechlin et al., 2003; Giese and Poggio, 2003; Botvinick, 2007; Badre, 2008; Fuster, 2004). Combining the Bayesian brain hypothesis with the hierarchical aspect of predictive coding provides a theoretical basis for computational mechanisms that drive a lifelong learning of the causal model of the world (Friston et al., 2014). Examples for how these different frameworks can be combined can be found in Yildiz and Kiebel (2011) and Yildiz et al. (2013).

As an example of a tight connection between prediction and sequences, one study investigating the electrophysiological responses in the song nucleus HVC of bengalese finch (Bouchard and Brainard, 2016) found evidence for an internal prediction of upcoming song syllables, based on sequential neuronal activity in HVC. As another example, a different study investigating single-cell recordings of neurons in the rat hippocampus found that sequences of neuronal activations during wheel-running between maze runs were predictive of the future behavior of the rats, including errors (Pastalkova et al., 2008). This finding falls in line with other studies showing that hippocampal sequences can correlate with future behavior (Pfeiffer, 2020).

1.4. What Are Sequences?

What does it mean to refer to neuronal activity as sequential? In the most common sense of the word, a sequence is usually understood as the serial succession of discrete elements or states. Likewise, when thinking of sequences, most people intuitively think of examples like "A, B, C,..." or "1, 2, 3,...." However, when extending this discrete concept to neuronal sequences, there are only few compelling examples where spike activity is readily interpretable as a discrete sequence, like the "domino-chain" activation observed in the birdbrain nucleus HVC (Hahnloser et al., 2002) (**Figure 1B**). As mentioned

before, we will use the word "sequence" to describe robust and reproducible spatiotemporal trajectories, which encode information to be processed or represented. Apart from the overwhelming body of literature reporting sequences in many different experimental settings (section 1.1), particularly interesting are the hippocampus (Bhalla, 2019; Pfeiffer, 2020) and entorhinal cortex (Zutshi et al., 2017; O'Neill et al., 2017). Due to the strong involvement of the hippocampus and the entorhinal cortex with sequences, the idea that neuronal sequences are also used in brain areas directly connected to them is not too far-fetched. For example, hippocampal-cortical interactions are characterized by sharp wave ripples (Buzsáki, 2015), which are effectively compressed spike sequences. Recent findings suggest that other cortical areas connected to the hippocampus use grid-cell like representations similar to space representation in the entorhinal cortex (Constantinescu et al., 2016; Stachenfeld et al., 2017). This is noteworthy because grid cells have been linked to sequence-like information processing (Zutshi et al., 2017; O'Neill et al., 2017). This suggests that at least areas connected to the hippocampus and entorhinal cortex are able to decode neuronal sequences.

The example of odor recognition shows that sequences are present even in circumstances where one intuitively would not expect them (**Figure 1C**). This very example does also show an interesting gap between a continuous and a discrete type of representation: The spatiotemporal trajectory is of a continuous nature, while the representation of the odor identity is characterized by discrete states and at a slower time scale. This gap also presents itself on another level. While we understand the term "neuronal sequence" to refer to a robust and reproducible spatiotemporal trajectory, in many cases these continuous state-space trajectories appear as a succession of quasi-discrete states (Abeles et al., 1995; Seidemann et al., 1996; Mazor and Laurent, 2005; Jones et al., 2007). In order to emphasize this interplay between continuous dynamics and discrete points we will denote such dynamics as *continuodiscrete* (see **Table 1**). In continuodiscrete dynamics, robust, and reproducible spatiotemporal trajectories are characterized by discrete points in state-space. As an example, in **Figure 1C** one can see the response of *in vivo* neurons in the gustatory cortex of rats, which is determined by the odor that is presented to the animal. The activity patterns of the neurons were analyzed with a hidden Markov model which revealed that the activity of the neuron ensemble can be described as a robust succession of discrete Markov states, where the system remains in a state for hundreds of milliseconds before quickly switching to another discrete state. These sequential visits to discrete states and the continuous expression of these states, specifically the switching between them, in terms of fast neuronal dynamics (here spiking neurons) is what we consider as continuodiscrete dynamics. Similar observations have been made in other experiments (Abeles et al., 1995; Seidemann et al., 1996; Mazor and Laurent, 2005; Rabinovich et al., 2001; Rivera et al., 2015) (see also **Figure 3**). The discrete states of a continuodiscrete sequence can be for example stable fixed points (Gros, 2009), or saddle points (Rabinovich et al., 2006, 2001) of the system, or simply points along a limit cycle trajectory (Yildiz and Kiebel, 2011; Yildiz et al., 2013), depending on the modeling approach (see section 2). Depending on the dynamical model, the system might leave a fixed point due to autonomously induced destabilization (Gros, 2007, 2009), noise (Rabinovich et al., 2006, 2001), or external input (Kurikawa and Kaneko, 2015; Toutounji and Pipa, 2014; Rivera et al., 2015; Hopfield, 1982).

Concepts similar to continuodiscrete trajectories have been introduced before. For example, in winner-less competition (WLC) (Rabinovich et al., 2000; Afraimovich et al., 2004b; Rabinovich et al., 2008), a system moves from one discrete metastable fixed-point (see **Table 1**) of the state space to the next, never settling for any state, similar to the fluctuations in a Lotka-Volterra system (Rabinovich et al., 2001) (see **Figure 3**). In winner-take-all (WTA) dynamics, like during memory recall in a Hopfield network (Hopfield, 1982), the system is attracted to one fixed point in which it will settle. Both WLC and WTA are thus examples of continuodiscrete dynamics. The concept of continuodiscrete dynamics also allows for dynamics which are characterized by an initial alteration between discrete states, before settling into a final state, as for example in Rivera et al. (2015). In section 2, we will look at different ways to model continuodiscrete neuronal dynamics.

For the brain, representing continuodiscrete trajectories seems to combine the best of two worlds: Firstly, the representation of discrete points forms the basis for the generalization and categorization of the sequence. For example, for the categorization of a specific movement sequence, it is not necessary to consider all the details of the sensory input, as it is sufficient to categorize the sequence type (dancing, walking, running) by recognizing the sequence of discrete points, as e.g., in Giese and Poggio (2003). Secondly, the brain requires a way of representing continuous dynamics to not miss important details. This is because key information can only be inferred by subtle variations within a sequence, as is often the case in our environment. For instance, when someone is talking, most of the speech content, i.e., what is being said, is represented by discrete points that describe a sequence of specific vocal tract postures. Additionally, there are subtle variations in the exact expression of these discrete points and the continuous dynamics connecting them, which let us infer about otherwise hidden states like the emotional state of the speaker (Birkholz et al., 2010; Kotz et al., 2003; Schmidt et al., 2006). Some of these subtle variations in the sensory input may be of importance to the brain, while others are not. For example, when listening to someone speaking, slight variations in the speaker's talking speed or pitch of voice might give hints about her mood, state of health, or hidden intentions. In other words, representing sensory input as continuodiscrete trajectories enables the recognition of invariances of the underlying movements without losing details.

There is growing evidence that sequences with discrete states like fixed points are a fundamental feature of cognitive and perceptual representations (e.g., Abeles et al., 1995; Seidemann et al., 1996; Mazor and Laurent, 2005; Jones et al., 2007). This feature may be at the heart of several findings in the cognitive sciences which suggest that human perception is chunked into discrete states, see VanRullen and Koch (2003) for some insightful examples. Assuming that the brain uses some form

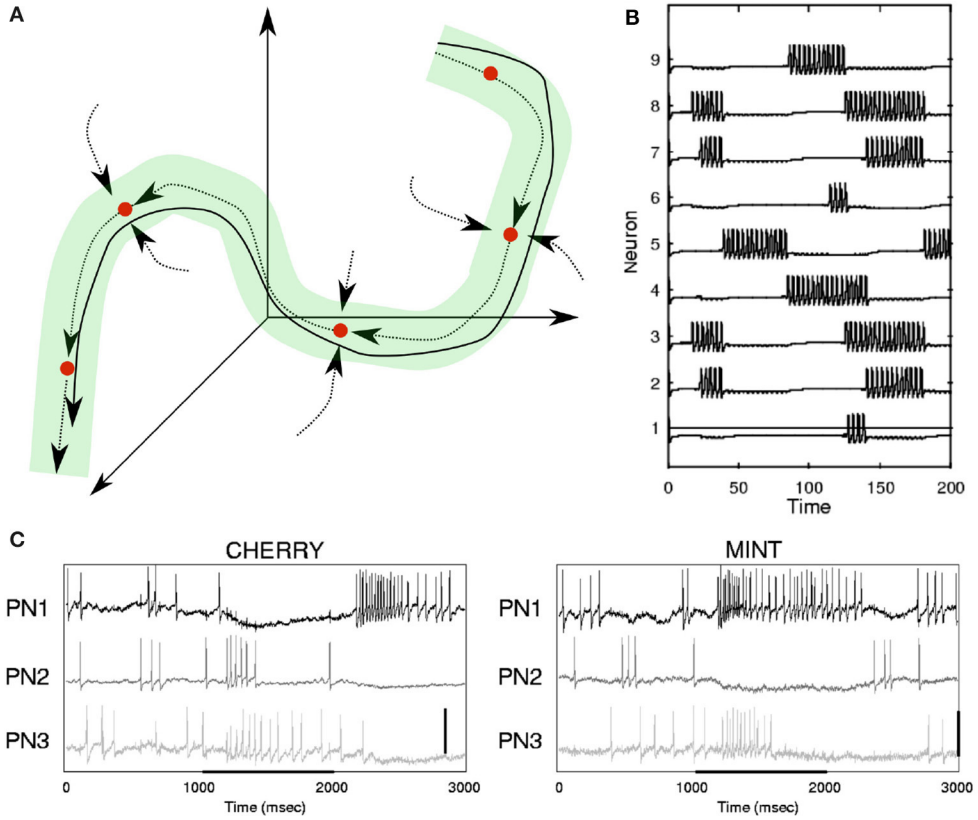


FIGURE 3 | (A) Illustration of continuodiscrete dynamics based on Stable Heteroclinic Channels (SHC, see section 2.2.2 and **Table 1**). The solid line represents a continuous heteroclinic trajectory in three-dimensional phase space and the dotted lines indicate invariant manifolds between saddle states (see **Table 1**). The green tube illustrates a Stable Heteroclinic Channel. All heteroclinic trajectories originating in the SHC will remain inside of it. This is a type of WLC dynamics. **(B)** Simulation of an SHC-trajectory based on Lotka-Volterra dynamics, where a point in phase space determines the firing rate of each neuron. **(C)** Neuronal responses to odor representation in the locust brain. **(B,C)** Are adapted from Rabinovich et al. (2001). Copyright (2001) by the American Physical Society.

of continuodiscrete dynamics to model sensory input, we will next consider neuronal sequence-generating mechanisms that may implement such dynamics and act as a generative model for recognition of sensory input. Importantly, as we are interested in generative models of sequential sensory input, we will only consider models that have the ability to autonomously generate sequential activity. Therefore, we are not going to discuss models where sequential activity is driven by sequential external input, as in models of non-autonomous neural networks (Toutounji and Pipa, 2014), or in models where intrinsic sequential neural activity is disrupted by bifurcation-inducing external input (Kurikawa and Kaneko, 2015).

2. NEURONAL NETWORK MODELS AS SEQUENCE GENERATORS

In order to explain sequential neuronal activity in networks of biological neurons, several models have been proposed, some of which we are going to review in the following sections. As this paper aims at a general overview of neuronal sequence-generating mechanisms and less at a detailed analysis, we will not cover the details and nuances of the presented dynamical models and refer the interested reader to the references given in the text.

2.1. Synfire Chains

Synfire chains are concatenated groups of excitatory neurons with convergent-divergent feed-forward connectivity, as illustrated in **Figure 4A** (Abeles, 1991; Diesmann et al., 1999). Synchronous activation of one group leads to the activation of the subsequent group in the chain after one synaptic delay (**Figure 4B**). It has been shown that the only stable operating mode in synfire chains is the synchronous mode where all neurons of a group spike in synchrony (Litvak et al., 2003). Synfire chains create sequences that are temporally highly precise (Abeles, 1991; Diesmann et al., 1999). Such temporally precise sequences have been observed in slices of the mouse primary visual cortex and in V1 of anaesthetized cats (Ikegaya et al., 2004), as well as in the HVC nucleus of the bird brain during song production (Hahnloser et al., 2002; Long et al., 2010), and in the frontal cortex of behaving monkeys (Prut et al., 1998; Abeles and Gat, 2001). While synfire chains make predictions that agree well with these observations, a striking mismatch between synfire chains and neuronal networks in the brain is the absence

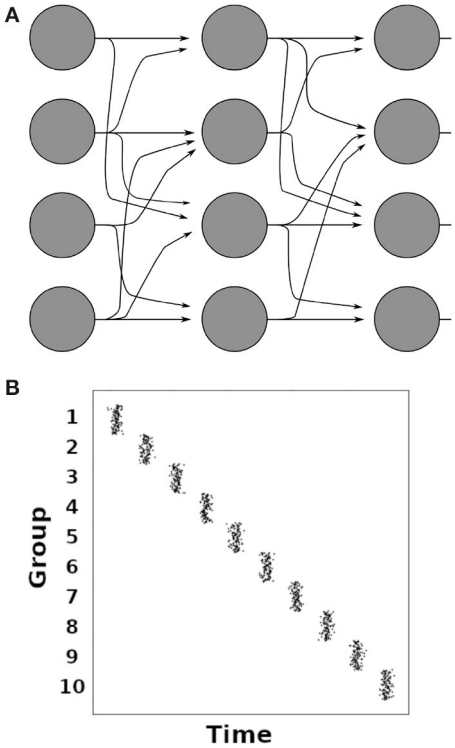


FIGURE 4 | (A) Illustration of a synfire chain between groups of neurons (filled circles). Arrows indicate excitatory connections. **(B)** Illustration of a spiking histogram of neurons in a synfire chain with 10 groups of 100 neurons each. The average time interval between the firing of two adjacent groups corresponds to one synaptic delay.

of recurrent connections in the synfire chain's feed-forward architecture. Modeling studies have shown that sequential activation similar to synfire chain activity can be achieved by changing a small fraction of the connections in a random neural network (Rajan et al., 2016; Chenkov et al., 2017), and that synfire chains can emerge in self-organizing recurrent neural networks under the influence of multiple interacting plasticity mechanisms (Zheng and Triesch, 2014). Such fractional changes of network connections were used to implement working memory (Rajan et al., 2016) or give a possible explanation for the occurrence of memory replay after one-shot learning (Chenkov et al., 2017). Such internally generated sequences have been proposed as a mechanism for memory consolidation, among other things (see Pezzulo et al., 2014 for a review).

2.2. Attractor Networks

2.2.1. Limit Cycles

Limit cycles are stable attractors in the phase space of a system, and they occur in practically every physical domain (Strogatz, 2018). A limit cycle is a closed trajectory, with fixed period and amplitude (**Figure 5**). Limit cycles occur frequently in biological and other dynamical systems, and the beating of the heart, or the periodic firing of a pacemaker neuron are examples of limit cycle behavior (Strogatz, 2018). They are of great interest to theoretical neuroscience, as periodic spiking activity can be represented by limit cycles, both on single-cell level (Izhikevich, 2007) and population level (Berry and Quoy, 2006; Jouffroy, 2007; Mi et al., 2017). They also play an important role in the emulation of human motion in robotics. While there are numerous ways to model human motion, one interesting approach is that of *dynamic motion primitives* (DMPs) (Schaal et al., 2007), which elegantly unifies the two different kinds of human motion, rhythmic and non-rhythmic motion, in one framework. The main idea of DMPs is that the limbs move as if they were pulled toward an attractor state. In the case of rhythmic motion, the attractor is given by a limit cycle, while in the case of motion strokes the attractor is a discrete point in space (Schaal et al., 2007). In Kiebel et al. (2009), Yildiz and Kiebel (2011), and Yildiz et al. (2013), the authors used a hierarchical generative model of sequence-generators based on limit cycles to model the generation and perception of birdsong and human speech.

2.2.2. Heteroclinic Trajectories

Another approach to modeling continuodiscrete dynamics are heteroclinic networks (Ashwin and Timme, 2005; Rabinovich et al., 2008) (see also **Table 1**). A heteroclinic network is a dynamical system with semi-stable states (saddle points) which are connected by invariant manifolds, so-called heteroclinic connections. Networks of coupled oscillators have been shown to give rise to phenomena like heteroclinic cycles (Ashwin and Swift, 1992; Ashwin et al., 2007). It has therefore been proposed that neuronal networks exhibit such heteroclinic behavior as well, which has been verified using simulations of networks of globally coupled Hodgkin-Huxley neurons (Hansel et al., 1993a,b; Ashwin and Borresen, 2004). Interestingly, heteroclinic networks can be harnessed to perform computational tasks (Ashwin and Borresen, 2005; Neves and Timme, 2012), and it has been shown that it is possible to implement any logic operation within such a network (Neves and Timme, 2012). Furthermore, the itinerancy in a heteroclinic network can be guided by external input, where the trajectory of fixed points discriminates between different inputs (Ashwin et al., 2007; Neves and Timme, 2012), which means that different inputs are encoded by different trajectories in phase space.

While theoretical neuroscience has progressed with research on heteroclinic behavior of coupled neural systems, concrete biological evidence is still sparse, as this requires a concrete and often complex mathematical model which is often beyond the more directly accessible research questions in biological science. Despite this, heteroclinic behavior has been shown to reproduce findings from single-cell recordings in insect olfaction (Rabinovich et al., 2001; Rivera et al., 2015) and olfactory bulb electroencephalography (EEG) in rabbits (Breakspear, 2001). Another study replicated the chaotic hunting behavior of a marine mollusk based on an anatomically plausible neuronal model with heteroclinic winnerless competition (WLC) dynamics (Varona et al., 2002), which is closely related to the dynamic alteration between states in a heteroclinic network (Rabinovich et al., 2000; Afraimovich et al., 2004b; Rabinovich et al., 2008). WLC was proposed as a general information processing principle for dynamical networks and is characterized

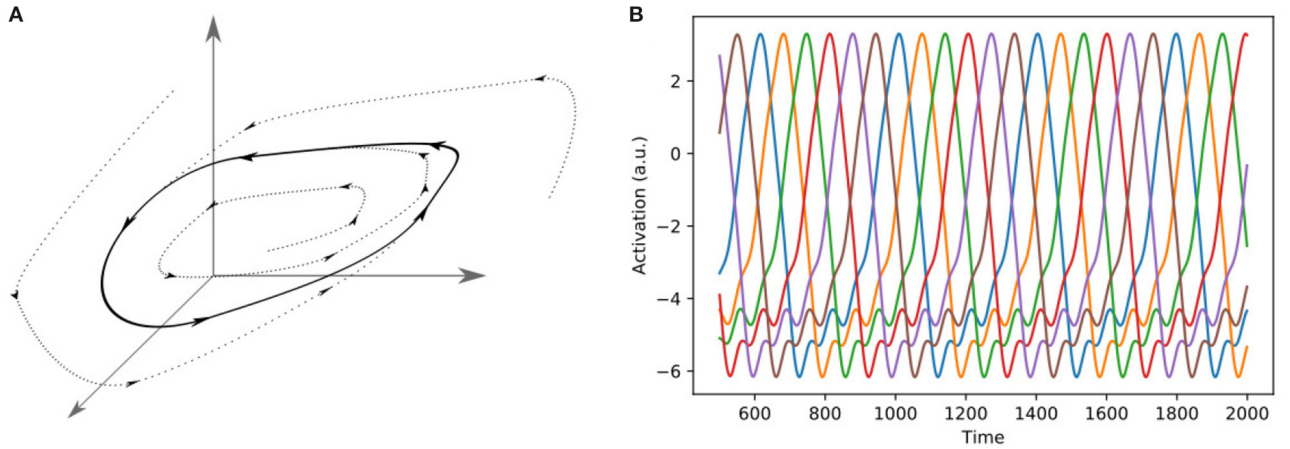


FIGURE 5 | Two different representations of a limit cycle. **(A)** A Limit cycle in three-dimensional phase space. In the case of a neuronal network, the dimensions of the phase space can be interpreted as the firing rates of the neurons. **(B)** Representation of a six-dimensional limit cycle as alternating activations of six different neurons.

by dynamic switching between network states, where the switching behavior is based on external input (Afraimovich et al., 2004b) (see **Table 1**). Importantly, the traveled trajectory identifies the received input, while any single state of the trajectory generally does not, see for example Neves and Timme (2012). In phase space representation, WLC can be achieved by open or closed sequences of heteroclinically concatenated saddle points. Such sequences are termed stable heteroclinic sequences (SHS) if the heteroclinic connections are dissipative, i.e., when a trajectory starting in a neighborhood close to the sequence remains close (Afraimovich et al., 2004a). While perturbations and external forcing can destroy stable heteroclinic sequences, it can be shown that even under such adverse circumstances, in many neurobiologically relevant situations the general sequential behavior of the system is preserved (Rabinovich et al., 2006). Such behavior is described by the concept of Stable Heteroclinic Channels (SHC) (see **Figure 3** and **Table 1**) (Rabinovich et al., 2006). A simple implementation of SHCs is based on the generalized Lotka-Volterra equations (Bick and Rabinovich, 2010; Rabinovich et al., 2001), which are a type of recurrent neural network implicitly implementing the WLC concept. The temporal precision of a system that evolves along an SHC is defined by the noise level as well as the eigenvalues of the invariant directions of the saddle points. Therefore, sequences along heteroclinic trajectories are reproducible although the exact timing of the sequence elements may be subject to fluctuation.

In a similar approach, recent theoretical work on the behavior of RNNs has introduced the concept of excitable network attractors, which are characterized by stable states of a system connected by excitable connections (Ceni et al., 2019). The conceptual idea of orbits between fixed points may further be implemented in different ways. For instance, transient activation of neuronal clusters can be achieved by autonomously driven destabilization of stable fixed points (Gros, 2007, 2009).

2.3. Hierarchical Sequence Generators

As briefly introduced in section 1.2, growing evidence suggests that the brain is organized into a hierarchy of different time scales, which enables the representation of different temporal features in its sensory input (e.g., Murray et al., 2014; Hasson et al., 2008; Cocchi et al., 2016; Mattar et al., 2016; Gauthier et al., 2012). Here the idea is that lower levels represent dynamics at faster time scales, which are integrated at higher levels that represent slower time scales. For example, speech consists of phonemes (fast time scales), which are integrated into increasingly slower representations of syllables, words, sentences, and a conversation (Hasson et al., 2008; Ding et al., 2016; Boemio et al., 2005). The combination of this hierarchical aspect of brain function with the Bayesian brain hypothesis and the concept of neuronal sequences suggests that the brain implicitly uses hierarchical continuodiscrete dynamical systems as generative models. One illustrative example of a hierarchical continuodiscrete process is given in **Figure 6**. In this example, the dynamics of the 2nd and 3rd level of the hierarchy are modeled by limit cycles and govern the evolution of parameters of the sequence-generating mechanisms at the levels below. Such an approach for a generative model for prediction and recognition of sensory data has been used to model birdsong and human speech recognition (Yildiz and Kiebel, 2011; Yildiz et al., 2013; Kiebel et al., 2009) (see **Figure 6**). In Yildiz and Kiebel (2011), the 3rd level represented sequential neuronal activity in area HVC (proper name, see also **Figure 1B**), and the 2nd level modeled activity in the robust nucleus of the arcopallium (RA). Similarly, in Rivera et al. (2015) the authors employed a hierarchical generative model with a heteroclinic sequence for a sequence-generating mechanism to model odor recognition in the insect brain. In a slightly different approach to hierarchical continuodiscrete modeling, hierarchical SHCs, implementing winnerless competition, were used to demonstrate how chunking of information can emerge, similar to memory representation in the brain (Fonollosa et al., 2015). One computational study

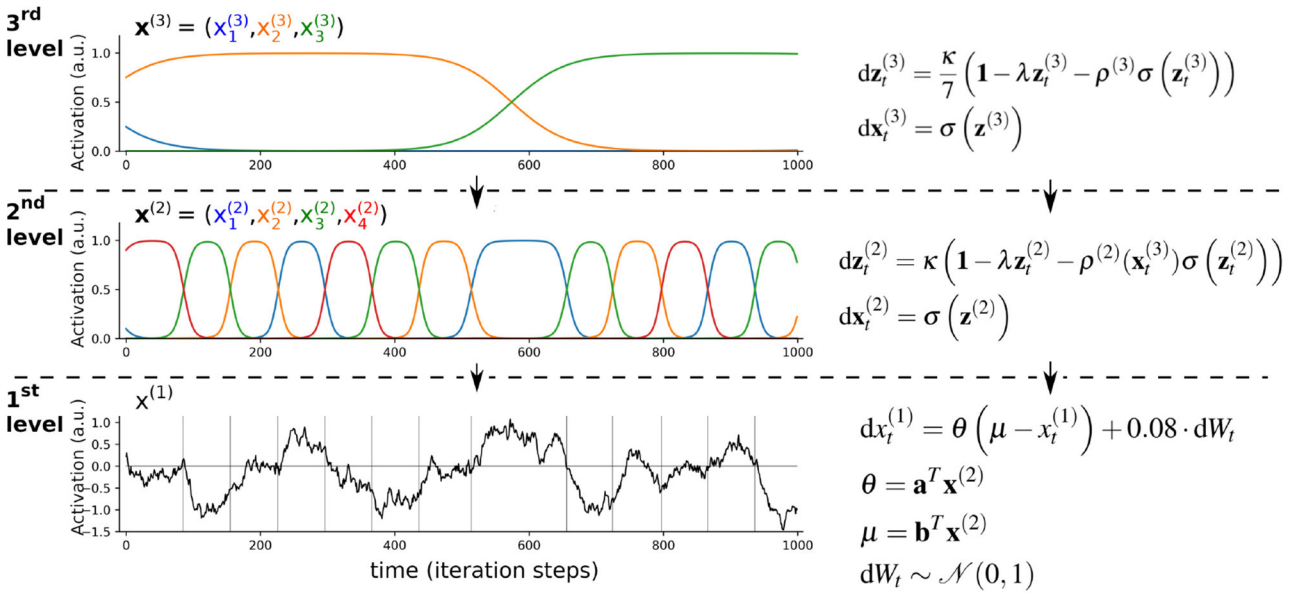


FIGURE 6 | Illustration of hierarchical continuodiscrete dynamics based on limit cycles. Slowly changing dynamics at the 3rd level parametrize the sequence of states of the faster changing 2nd-level dynamics $z^{(2)}$. As the dynamics of variables $\vec{x}^{(2)}$ and $\vec{x}^{(3)}$ change between the states "on" and "off," their behavior constitutes continuodiscrete WLC dynamics. At around iteration step 600, the green unit at the 3rd level (element $\vec{x}_3^{(3)}$) becomes active, which changes the 2nd-level sequential dynamics from red→green→orange→blue→red to green→orange→red→blue→green. This is achieved by a change of the 2nd-level connectivity matrix $\rho^{(2)}$ which depends on the 3rd-level variable $\vec{x}^{(3)}$. In this toy example, the 2nd-level dynamics model the evolution of the parameters of an Ornstein-Uhlenbeck process (black graph showing the evolution of variable $x^{(1)}$). In the framework of hierarchical generative modeling, the 1st level would correspond to an agent's predictions of its sensory input, while the higher levels are the hidden states of the agent's generative model. This hierarchical parametrization of sequences is similar to the approach in Kiebel et al. (2009). The dot product between vectors $\boldsymbol{b} = (0.6, 0, -1, -0.3)^T$ and $\vec{x}^{(2)}$ determines the 1st-level attractor μ. The rate parameter Θ is parametrized by vector $\boldsymbol{a} = (1, 0.5, 1.2, 0.8)^T$ and its dot product with $\vec{x}^{(2)}$. $\sigma(\cdot)$ is the softmax function which is applied element-wise. **1** denotes a vector of ones. $\kappa = 2$, $\lambda = 1/8$. Gray vertical lines in the 1st level mark the time-points where states in the 2nd level change. This hierarchical parametrization of sequences is similar to the approach in Kiebel et al. (2009). Similar hierarchical autonomous models can be used as a generative model for Bayesian inference to achieve prediction and recognition of sequential data, as has for example been done in Yildiz and Kiebel (2011) and Yildiz et al. (2013).

provided a proof of principle that complex behavior, like handwriting, can be decomposed into a hierarchical organization of stereotyped dynamical flows on manifolds of lower dimensions (Perdikis et al., 2011). These stereotyped dynamics can be regarded as the discrete points in a continuodiscrete sequence, which gave rise to complex and flexible behavior.

In the following section, we will briefly review how sequential methods have been used for problems in neuroscience and especially AI. Afterwards, we will review evidence for the organization of neuronal sequences into a hierarchy of time scales.

3. RECOGNITION OF SEQUENCES

Although neuronal sequence models, such as the ones introduced in the preceding sections have been used to explain experimentally observed neuronal activity, these models by themselves do not explain how predictions are formed about the future trajectory of a sequence. To take the example of song production and recognition in songbirds, a sequence-generating model of birdsong generation is not sufficient to model or explain how a listening bird recognizes a song (Yildiz and Kiebel, 2011). Given a generative model, recognition of a song corresponds to statistical model inversion (Watzenig, 2007; Ulrych et al., 2001). A simple example of such a scheme is provided in Bitzer and Kiebel (2012), where RNNs are used as a generative model such that model inversion provides for an online recognition model. As shown in Friston et al. (2011), one can also place such a generative model into the active inference framework to derive a model that not only recognizes sequential movements from visual input but also generates continuodiscrete movement patterns. Generative models are not only interesting from a cognitive neuroscience perspective but also point at a shared interest with the field of artificial intelligence and specifically machine learning, to find a mechanistic understanding of how spatiotemporally structured sensory input can be recognized by an artificial or a biological agent. In the following, we will discuss how both fields seem to converge on the conceptual idea that generative models should be spatiotemporally structured and hierarchical.

3.1. Sequence Recognition in Machine Learning

The most widely-used models for discrete sequence generation are hidden Markov models (HMM) and their time-dependent generalisation, hidden semi-Markov models (HSMM) (Yu,

2015). In particular, HMMs and HSMMs are standard tools in a wide range of applications concerned with e.g., speech recognition (Liu et al., 2018; Zen et al., 2004; Deng et al., 2006) and activity recognition (Duong et al., 2005). Furthermore, they have often been used for the analysis of neuronal activity (Tokdar et al., 2010) and human behavior in general (Eldar et al., 2011). Similar to HSMMs, artificial RNNs are used in machine learning for classifying and predicting time series data. When training a generic RNN for prediction and classification of time series data, one faces various challenges, most notably incorporating information about long-term dependencies in the data. To address these dependencies, specific RNN architectures have been proposed, such as *long-short term memory* (LSTM) networks (Gers et al., 1999) and *gate recurrent units* (GRU) (Chung et al., 2014). In a common LSTM network, additionally to the output variable, the network computes an internal memory variable. This endows the network with high flexibility. LSTM networks belong to the most successful and most widely applied RNN architectures, with applications in virtually every field involving time-series data, or any data structure with long-range dependencies (Yu et al., 2019; LeCun et al., 2015). Another RNN approach is *reservoir computing* (RC), which started with the development of echo-state networks and liquid state machines in the early 2000s (Lukoševičius et al., 2012; Jaeger, 2001; Maass et al., 2002). In RC, sequential input is fed to one or more input neurons. Those neurons are connected with a *reservoir* of randomly connected neurons, which in turn are connected to one or more output neurons. Connections in the reservoir are pseudo-randomized to elicit dynamics at the edge of chaos (Yildiz et al., 2012), leading to a spatiotemporal network response in the form of reverberations over multiple time scales. RC networks have successfully been applied in almost every field of machine learning and data science, such as speech recognition, handwriting recognition, robot motor control, and financial forecasting (Lukoševičius et al., 2012; Tanaka et al., 2019).

While there is a lot of research on neurobiologically plausible learning paradigms for RNNs (Sussillo and Abbott, 2009; Miconi, 2017; Taherkhani et al., 2020), one possible approach for understanding the role of neuronal sequences is to use neurobiologically more plausible sequence generation models, which can act as generative models of the causal dynamic relationships in the environment. A natural application would be the development of recognition models based on Bayesian inference (Bitzer and Kiebel, 2012), and more specifically in terms of variational inference (Friston et al., 2006; Daunizeau et al., 2009).

3.2. Biological and Artificial Inferential Hierarchies

In neuroscience and the cognitive sciences, the brain is often viewed as a hierarchical system, where a functional hierarchy can be mapped to the structural hierarchy of the cortex (Badre, 2008; Koechlin et al., 2003; Kiebel et al., 2008). The best example of such a hierarchical organization is the visual system, for which the existence of both a functional and an equivalent structural hierarchy is established (Felleman and Van Essen, 1991). Cells in lower levels of the hierarchy encode simple features and have smaller receptive fields than cells further up the hierarchy, which posses larger receptive fields and encode more complex patterns by integrating information from lower levels (Hubel and Wiesel, 1959; Zeki and Shipp, 1988; Giese and Poggio, 2003). This functional hierarchy is mediated by an asymmetry of recurrent connectivity in the visual stream, where forward connections to higher layers are commonly found to have fast, excitatory effects on the post-synaptic neurons, while feedback connections act in a slower, modulatory manner (Zeki and Shipp, 1988; Sherman and Guillery, 1998). Moreover, neuroimaging studies have shown that the brain is generally organized into a modular hierarchical structure (Bassett et al., 2010; Meunier et al., 2009, 2010). This is substantiated by other network-theoretical characteristics of the brain, like its scale-free property (Eguiluz et al., 2005), which is a natural consequence of modular hierarchy (Ravasz and Barabási, 2003). Hierarchies also play an important role in cognitive neuroscience as most if not all types of behavior, as well as cognitive processes, can be described in a hierarchical fashion. For example, making a cup of tea can be considered a high-order goal in a hierarchy with subgoals that are less abstract and temporally less extended. In the example of making a cup of tea, these subgoals can be: (i) putting a teabag into a pot, (ii) pouring hot water into the pot, and (iii) pouring tea into a cup (example adopted from Botvinick, 2007).

3.2.1. A Hierarchy of Time Scales

Importantly, all theories of cortical hierarchies of function share the common assumption that primary sensory regions encode rather quickly changing dynamics representing the fast features of sensory input, and that those regions are at the bottom of the hierarchy, while temporally more extended or more abstract representations are located in higher order cortices. This principle has been conceptualized as a "hierarchy of time scales" (Kiebel et al., 2008; Hasson et al., 2008; Koechlin et al., 2003; Badre, 2008; Kaplan et al., 2020). In this view, levels further up the hierarchy code for more general characteristics of the environment and inner cognitive processes, which generally change slowly (Hasson et al., 2008; Koechlin et al., 2003; Badre, 2008). For example, although the visual hierarchy is typically understood as a spatial hierarchy, experimental evidence is emerging that it is also a hierarchy of time scales (Cocchi et al., 2016; Gauthier et al., 2012; Mattar et al., 2016). Importantly, the information exchange in such a hierarchy is bidirectional. While top-down information can be regarded as the actions of a generative model trying to predict the sensory input (Dayan et al., 1995; Friston, 2005), recognition is achieved by bottom-up information that provides higher levels in the hierarchy with information about the sensory input, see also Yildiz and Kiebel (2011) and Yildiz et al. (2013) for illustrations of this concept. A related finding is an experimentally observed hierarchy of time scales with respect to the time lag of the autocorrelation of neuronal measurements (e.g., Murray et al., 2014). Here, it was found that the decay of autocorrelation was fastest for sensory areas ($<$100 ms) but longest for prefrontal areas like ACC ($>$300 ms).

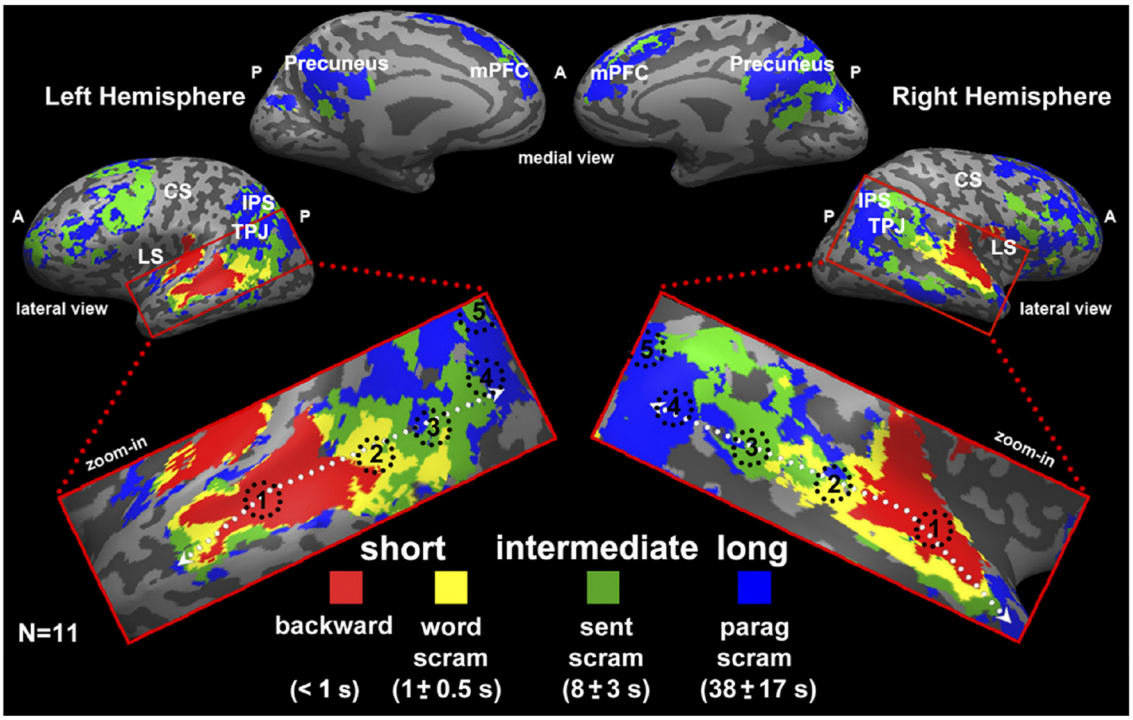


FIGURE 7 | Study by Lerner et al. (2011) as an example for representations in a hierarchy of time scales. Here, the authors used fMRI and a between-subject correlational analysis to categorize brain voxels according to four levels of representation. These four levels were fast dynamics of auditory input (red), words (yellow), sentences (green), and paragraphs (blue). Results are displayed on a so-called inflated cortical surface. Figure reprinted from Lerner et al. (2011).

The importance of cognition based on spatiotemporal structure at multiple time scales is also illustrated by various computational modeling studies. In one study, robots were endowed with a neural network whose parameters were let free to evolve over time to optimize performance during a navigation task (Nolfi, 2002). After some time, the robots had evolved neural assemblies with representations at clearly distinct time scales: one assembly had assumed a quickly changing, short time scale associated with immediate sensory input while another assembly had adopted a long time scale, associated with an integration of information over an extended period of time, which was necessary for succeeding at the task. Another modeling study showed that robots with neuronal populations of strongly differing time-constants performed their tasks significantly better than when endowed only with units of approximately identical time-constants (Yamashita and Tani, 2008). In Botvinick (2007) it was shown that, after learning, a neural network with a structural hierarchy similar to the one proposed for the frontal cortex had organized in such a way that high-level units coded for temporal context while low-level units encoded fast responses similar to the role assigned to sensory and motor regions in theories of hierarchical cortical processing (Kiebel et al., 2008; Alexander and Brown, 2018; Rao and Ballard, 1999; Botvinick, 2008; Badre, 2008; Koechlin et al., 2003; Fuster, 2004).

The principle of representing spatiotemporal dynamics at multiple time scales has also been used to model birdsong generation and inference in songbirds by combining a hierarchically structured RNN with a model of songbirds' vocal tract dynamics (Yildiz and Kiebel, 2011). The system consisted of three levels, each of which was governed by the sequential dynamics of an RNN following a limit cycle. The sequential dynamics were influenced both by top-down predictions, and bottom-up prediction errors. In another study, the same concept was applied to the recognition of human speech (Yildiz et al., 2013). The resulting inference scheme was able to recognize spoken words, even under adversarial circumstances like accelerated speech, since it inferred and adapted parameters in an online fashion during the recognition process. The same principle can also be translated to very different types of input, see Rivera et al. (2015) for an example of insect olfaction.

3.2.2. A Hierarchy of Time Scales: Neuroimaging Evidence

Experimental evidence for the hypothesis of a hierarchy of time scales has been reported in several neuroimaging studies (Koechlin et al., 2003; Hasson et al., 2008; Lerner et al., 2011; Gauthier et al., 2012; Cocchi et al., 2016; Mattar et al., 2016; Baldassano et al., 2017; Gao et al., 2020), two of which we are going to briefly discuss in the following. One functional magnetic resonance imaging (fMRI) study investigated the temporal receptive windows (TRW) of several brain regions in the human brain (Hasson et al., 2008). The TRW of an area is the time-interval over which the region "integrates" incoming information, in order to extract meaning over a specific temporal

scale. It was found that regions, such as the primary visual cortex exhibited rather short TRW, while high order regions exhibited intermediate to long TRW (Hasson et al., 2008). Similarly, in Lerner et al. (2011) the same principle was tested with temporally structured auditory input, i.e., speech. Using fMRI, the authors found evidence for a hierarchy of time scales in specific brain areas. The different time scales represented fast auditory input, words, sentences and paragraphs (see **Figure 7**).

3.2.3. A Hierarchy of Time Scales: Machine Learning

Not surprisingly, the importance of hierarchies of time scales is well-established within the machine learning community (El Hihi and Bengio, 1996; Malhotra et al., 2015). Current state-of-the-art RNN architectures used for prediction and classification of complex time series data are based on recurrent network units organized as temporal hierarchies. Notable examples are the clockwork RNN (Koutnik et al., 2014), gated feedback RNN (Chung et al., 2015), hierarchical multi-scale RNN (Chung et al., 2016), fast-slow RNN (Mujika et al., 2017), and higher order RNNs (HORNNs) (Soltani and Jiang, 2016). These modern RNN architectures have found various applications in motion classification (Neverova et al., 2016; Yan et al., 2018), speech synthesis (Wu and King, 2016; Achanta and Gangashetty, 2017; Zhang and Woodland, 2018), recognition (Chan et al., 2016), and other related areas (Liu et al., 2015; Krause et al., 2017; Kurata et al., 2017). These applications of hierarchical RNN architectures further confirm the relevance of hierarchically organized sequence generators for capturing complex dynamics in our everyday environments.

4. CONCLUSION

Here, we have reviewed the evidence that our brain senses its environment as sequential sensory input, and consequently, uses neuronal sequences for predicting future sensory input. Although the general idea that the brain is a prediction device has by now become a mainstream guiding principle in cognitive neuroscience, it is much less clear how exactly the brain computes these predictions. We have reviewed results from different areas of the neurosciences that the brain may achieve this by using a hierarchy of time scales, specifically a hierarchy of sequential dynamics. If this were the case, the question would be whether already known neuroscience results in specific areas can be re-interpreted as evidence for the brain's operations in such a hierarchy of time scales. Such an interpretation is quite natural for neuroscience fields like auditory processing, where such a temporal hierarchy is most evident. But it is much less evident for other areas, like for example decision-making. To further test this suggested theory of brain function, researchers need to design experimental paradigms which are specifically geared toward testing what probabilistic inference mechanisms the brain uses to predict its input at different time scales, and select its own actions. Importantly, hierarchical computational modeling approaches as reviewed here could be used to further provide theoretical evidence of the underlying multi-scale inference mechanism and generate new predictions that can be tested experimentally.

What we found telling is that recent advances in machine learning converge on similar ideas of representing multi scale dynamics in sensory data, although with a different motivation and different aims. The simple reason for this convergence may be that much of the sensory data that is input to machine learning implementations is similar to the kind of sensory input experienced by humans, as for example in videos and speech data. Therefore, we believe that as computational modeling in the neurosciences as reviewed here will gain traction, there will be useful translations form the neurosciences to machine learning applications.

AUTHOR CONTRIBUTIONS

DM and SK contributed to the conception of the manuscript. SF wrote the manuscript, with contributions by DM and SK. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft), SFB 940/2 - Project ID 178833530 A9, TRR 265/1 - Project ID 402170461 B09, and as part of Germany's Excellence Strategy - EXC 2050/1 - Project ID 390696704 -Cluster of Excellence Centre for Tactile Internet with Human-in-the-Loop (CeTI) of Technische Universität Dresden.

REFERENCES

Abeles, M. (1991). *Corticonics: Neural Circuits of the Cerebral Cortex*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511574566

Abeles, M., Bergman, H., Gat, I., Meilijson, I., Seidemann, E., Tishby, N., et al. (1995). Cortical activity flips among quasi-stationary states. *Proc. Natl. Acad. Sci. U.S.A.* 92, 8616–8620. doi: 10.1073/pnas.92.19.8616

Abeles, M., and Gat, I. (2001). Detecting precise firing sequences in experimental data. *J. Neurosci. Methods* 107, 141–154. doi: 10.1016/S0165-0270(01)00364-8

Achanta, S., and Gangashetty, S. V. (2017). Deep elman recurrent neural networks for statistical parametric speech synthesis. *Speech Commun.* 93, 31–42. doi: 10.1016/j.specom.2017.08.003

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., and Friston, K. J. (2013). The computational anatomy of psychosis. *Front. Psychiatry* 4:47. doi: 10.3389/fpsyt.2013.00047

Afraimovich, V., Zhigulin, V., and Rabinovich, M. (2004a). On the origin of reproducible sequential activity in neural circuits. *Chaos* 14, 1123–1129. doi: 10.1063/1.1819625

Afraimovich, V. S., Rabinovich, M. I., and Varona, P. (2004b). Heteroclinic contours in neural ensembles and the winnerless competition principle. *Int. J. Bifurc. Chaos* 14, 1195–1208. doi: 10.1142/S0218127404009806

Aitchison, L., and Lengyel, M. (2017). With or without you: predictive coding and bayesian inference in the brain. *Curr. Opin. Neurobiol.* 46, 219–227. doi: 10.1016/j.conb.2017.08.010

Alexander, W. H., and Brown, J. W. (2018). Frontal cortex function as derived from hierarchical predictive coding. *Sci. Rep*. 8:3843. doi: 10.1038/s41598-018-21407-9

Ashwin, P., and Borresen, J. (2004). Encoding via conjugate symmetries of slow oscillations for globally coupled oscillators. *Phys. Rev. E* 70:026203. doi: 10.1103/PhysRevE.70.026203

Ashwin, P., and Borresen, J. (2005). Discrete computation using a perturbed heteroclinic network. *Phys. Lett. A* 347, 208–214. doi: 10.1016/j.physleta.2005.08.013

Ashwin, P., Orosz, G., Wordsworth, J., and Townley, S. (2007). Dynamics on networks of cluster states for globally coupled phase oscillators. *SIAM J. Appl. Dyn. Syst*. 6, 728–758. doi: 10.1137/070683969

Ashwin, P., and Swift, J. W. (1992). The dynamics of n weakly coupled identical oscillators. *J. Nonlin. Sci*. 2, 69–108. doi: 10.1007/BF02429852

Ashwin, P., and Timme, M. (2005). Nonlinear dynamics: when instability makes sense. *Nature* 436:36. doi: 10.1038/436036b

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends Cogn. Sci*. 12, 193–200. doi: 10.1016/j.tics.2008.02.004

Baeg, E., Kim, Y., Huh, K., Mook-Jung, I., Kim, H., and Jung, M. (2003). Dynamics of population code for working memory in the prefrontal cortex. *Neuron* 40, 177–188. doi: 10.1016/S0896-6273(03)00597-X

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron* 95, 709–721. doi: 10.1016/j.neuron.2017.06.041

Bassett, D. S., Greenfield, D. L., Meyer-Lindenberg, A., Weinberger, D. R., Moore, S. W., and Bullmore, E. T. (2010). Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Comput. Biol*. 6:e1000748. doi: 10.1371/journal.pcbi.1000748

Bazhenov, M., Stopfer, M., Rabinovich, M., Abarbanel, H. D., Sejnowski, T. J., and Laurent, G. (2001). Model of cellular and network mechanisms for odor-evoked temporal patterning in the locust antennal lobe. *Neuron* 30, 569–581. doi: 10.1016/S0896-6273(01)00286-0

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., et al. (2008). Probabilistic population codes for bayesian decision making. *Neuron* 60, 1142–1152. doi: 10.1016/j.neuron.2008.09.021

Berry, H., and Quoy, M. (2006). Structure and dynamics of random recurrent neural networks. *Adapt. Behav*. 14, 129–137. doi: 10.1177/105971230601400204

Bhalla, U. S. (2019). Dendrites, deep learning, and sequences in the hippocampus. *Hippocampus* 29, 239–251. doi: 10.1002/hipo.22806

Bick, C., and Rabinovich, M. I. (2010). On the occurrence of stable heteroclinic channels in lotka-volterra models. *Dyn. Syst*. 25, 97–110. doi: 10.1080/14689360903322227

Birkholz, P., Kroger, B. J., and Neuschaefer-Rube, C. (2010). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Trans. Audio Speech Lang. Process*. 19, 1422–1433. doi: 10.1109/TASL.2010.2091632

Bitzer, S., and Kiebel, S. J. (2012). Recognizing recurrent neural networks (RRNN): Bayesian inference for recurrent neural networks. *Biol. Cybernet*. 106, 201–217. doi: 10.1007/s00422-012-0490-x

Boemio, A., Fromm, S., Braun, A., and Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat. Neurosci*. 8, 389–395. doi: 10.1038/nn1409

Botvinick, M. M. (2007). Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. *Philos. Trans. R. Soc. B Biol. Sci*. 362, 1615–1626. doi: 10.1098/rstb.2007.2056

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci*. 12, 201–208. doi: 10.1016/j.tics.2008.02.009

Bouchard, K. E., and Brainard, M. S. (2016). Auditory-induced neural dynamics in sensory-motor circuitry predict learned temporal and sequential statistics of birdsong. *Proc. Natl. Acad. Sci. U.S.A*. 113, 9641–9646. doi: 10.1073/pnas.1606725113

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *J. Gen. Psychol*. 49, 229–240. doi: 10.1080/00221309.1953.9710088

Branco, T., Clark, B. A., and Häusser, M. (2010). Dendritic discrimination of temporal input sequences in cortical neurons. *Science* 329, 1671–1675. doi: 10.1126/science.1189664

Breakspear, M. (2001). Perception of odors by a nonlinear model of the olfactory bulb. *Int. J. Neural Syst*. 11, 101–124. doi: 10.1142/S0129065701000564

Buonomano, D. V., and Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci*. 10:113. doi: 10.1038/nrn2558

Buzsáki, G. (2015). Hippocampal sharp wave-ripple: a cognitive biomarker for episodic memory and planning. *Hippocampus* 25, 1073–1188. doi: 10.1002/hipo.22488

Ceni, A., Ashwin, P., and Livi, L. (2019). Interpreting recurrent neural networks behaviour via excitable network attractors. *Cogn. Comput*. 12, 330–356. doi: 10.1007/s12559-019-09634-2

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Shanghai: IEEE), 4960–4964. doi: 10.1109/ICASSP.2016.7472621

Chenkov, N., Sprekeler, H., and Kempter, R. (2017). Memory replay in balanced recurrent networks. *PLoS Comput. Biol*. 13:e1005359. doi: 10.1371/journal.pcbi.1005359

Chung, J., Ahn, S., and Bengio, Y. (2016). Hierarchical multiscale recurrent neural networks. *arXiv* arXiv:1609.01704.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* 1412.3555.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2015). "Gated feedback recurrent neural networks," in *International Conference on Machine Learning* (Lille), 2067–2075.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci*. 36, 181–204. doi: 10.1017/S0140525X12000477

Cocchi, L., Sale, M. V., Gollo, L. L., Bell, P. T., Nguyen, V. T., Zalesky, A., et al. (2016). A hierarchy of timescales explains distinct effects of local inhibition of primary visual cortex and frontal eye fields. *Elife* 5:e15252. doi: 10.7554/eLife.15252

Colombo, M., and Seriès, P. (2012). Bayes in the brain—on bayesian modelling in neuroscience. *Br. J. Philos. Sci*. 63, 697–723. doi: 10.1093/bjps/axr043

Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science* 352, 1464–1468. doi: 10.1126/science.aaf0941

Crowe, D. A., Averbeck, B. B., and Chafee, M. V. (2010). Rapid sequences of population activity patterns dynamically encode task-critical spatial information in parietal cortex. *J. Neurosci*. 30, 11640–11653. doi: 10.1523/JNEUROSCI.0954-10.2010

Daunizeau, J., Friston, K. J., and Kiebel, S. J. (2009). Variational bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Phys. D* 238, 2089–2118. doi: 10.1016/j.physd.2009.08.002

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz machine. *Neural Comput*. 7, 889–904. doi: 10.1162/neco.1995.7.5.889

Deneve, S. (2008). Bayesian spiking neurons I: inference. *Neural Comput*. 20, 91–117. doi: 10.1162/neco.2008.20.1.91

Deng, L., Yu, D., and Acero, A. (2006). Structured speech modeling. *IEEE Trans. Audio Speech Lang. Process*. 14, 1492–1504. doi: 10.1109/TASL.2006.878265

Dezfouli, A., Lingawi, N. W., and Balleine, B. W. (2014). Habits as action sequences: hierarchical action control and changes in outcome value. *Philos. Trans. R. Soc. B Biol. Sci*. 369:20130482. doi: 10.1098/rstb.2013.0482

Diesmann, M., Gewaltig, M. O., and Aertsen, A. (1999). Stable propagation of synchronous spiking in cortical neural networks. *Nature* 402:529. doi: 10.1038/990101

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nat. Neurosci*. 19, 158–164. doi: 10.1038/nn.4186

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262042383.001.0001

Dragoi, G., and Tonegawa, S. (2011). Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature* 469:397. doi: 10.1038/nature09633

Duong, T. V., Bui, H. H., Phung, D. Q., and Venkatesh, S. (2005). "Activity recognition and abnormality detection with the switching hidden semi-Markov model," in *2005 IEEE Computer Society Conference on Computer Vision and*

Pattern Recognition (CVPR'05), Vol. 1 (San Diego, CA: IEEE), 838–845. doi: 10.1109/CVPR.2005.61

Eguiluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M., and Apkarian, A. V. (2005). Scale-free brain functional networks. *Phys. Rev. Lett*. 94:018102. doi: 10.1103/PhysRevLett.94.018102

El Hihi, S., and Bengio, Y. (1996). "Hierarchical recurrent neural networks for long-term dependencies," in *Advances in Neural Information Processing Systems*, 493–499.

Eldar, E., Morris, G., and Niv, Y. (2011). The effects of motivation on response rate: a hidden semi-Markov model analysis of behavioral dynamics. *J. Neurosci. Methods* 201, 251–261. doi: 10.1016/j.jneumeth.2011.06.028

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415:429. doi: 10.1038/415429a

Feinberg, E. A., and Shwartz, A. (2012). *Handbook of Markov Decision Processes: Methods and Applications*, Vol. 40. Boston, MA: Springer Science & Business Media.

Feldman, J. (2001). Bayesian contour integration. *Percept. Psychophys*. 63, 1171–1182. doi: 10.3758/BF03194532

Felleman, D. J., and Van Essen, D. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

FitzGerald, T. H., Hämmerer, D., Friston, K. J., Li, S. C., and Dolan, R. J. (2017). Sequential inference as a mode of cognition and its correlates in fronto-parietal and hippocampal brain regions. *PLoS Comput. Biol*. 13:e1005418. doi: 10.1371/journal.pcbi.1005418

Fletcher, P. C., and Frith, C. D. (2009). Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci*. 10, 48–58. doi: 10.1038/nrn2536

Fonollosa, J., Neftci, E., and Rabinovich, M. (2015). Learning of chunking sequences in cognition and behavior. *PLoS Comput. Biol*. 11:e1004592. doi: 10.1371/journal.pcbi.1004592

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci*. 360, 815–836. doi: 10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci*. 11, 127–138. doi: 10.1038/nrn2787

Friston, K., and Buzsáki, G. (2016). The functional anatomy of time: what and when in the brain. *Trends Cogn. Sci*. 20, 500–511. doi: 10.1016/j.tics.2016.05.001

Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. B Biol. Sci*. 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol*. 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Friston, K., Mattout, J., and Kilner, J. (2011). Action understanding and active inference. *Biol. Cybernet*. 104, 137–160. doi: 10.1007/s00422-011-0424-z

Friston, K. J., Stephan, K. E., Montague, R., and Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *Lancet Psychiatry* 1, 148–158. doi: 10.1016/S2215-0366(14)70275-5

Fuster, J. M. (2004). Upper processing stages of the perception-action cycle. *Trends Cogn. Sci*. 8, 143–145. doi: 10.1016/j.tics.2004.02.004

Gao, R., van den Brink, R. L., Pfeffer, T., and Voytek, B. (2020). Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture. *Elife* 9:e61277. doi: 10.7554/eLife.61277

Gauthier, B., Eger, E., Hesselmann, G., Giraud, A. L., and Kleinschmidt, A. (2012). Temporal tuning properties along the human ventral visual stream. *J. Neurosci*. 32, 14433–14441. doi: 10.1523/JNEUROSCI.2467-12.2012

Gelman, A., Simpson, D., and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy* 19:555. doi: 10.3390/e19100555

Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). *Learning to Forget: Continual Prediction With LSTM*. Stevenage: Institution of Engineering and Technology. doi: 10.1049/cp:19991218

Giese, M. A., and Poggio, T. (2003). Cognitive neuroscience: neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci*. 4:179. doi: 10.1038/nrn1057

Giraud, A. L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci*. 15:511. doi: 10.1038/nn.3063

Gros, C. (2007). Neural networks with transient state dynamics. *New J. Phys*. 9:109. doi: 10.1088/1367-2630/9/4/109

Gros, C. (2009). Cognitive computation with autonomously active neural networks: an emerging field. *Cogn. Comput*. 1, 77–90. doi: 10.1007/s12559-008-9000-9

Hahnloser, R. H., Kozhevnikov, A. A., and Fee, M. S. (2002). An ultra-sparse code underliesthe generation of neural sequences in a songbird. *Nature* 419:65. doi: 10.1038/nature00974

Hansel, D., Mato, G., and Meunier, C. (1993a). Clustering and slow switching in globally coupled phase oscillators. *Phys. Rev. E* 48:3470. doi: 10.1103/PhysRevE.48.3470

Hansel, D., Mato, G., and Meunier, C. (1993b). Phase dynamics for weakly coupled hodgkin-huxley neurons. *Europhys. Lett*. 23:367. doi: 10.1209/0295-5075/23/5/011

Harvey, C. D., Coen, P., and Tank, D. W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* 484:62. doi: 10.1038/nature10918

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *J. Neurosci*. 28, 2539–2550. doi: 10.1523/JNEUROSCI.5487-07.2008

Hawkins, J., George, D., and Niemasik, J. (2009). Sequence memory for prediction, inference and behaviour. *Philos. Trans. R. Soc. B Biol. Sci*. 364, 1203–1209. doi: 10.1098/rstb.2008.0322

Helmholtz, H. V. (1867). *Handbuch der Physiologischen Optik*. Leipzig: Voss.

Hinton, G. E., and Sejnowski, T. J. (1983). "Optimal perceptual inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 448 (New York, NY: Citeseer).

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A*. 79, 2554–2558. doi: 10.1073/pnas.79.8.2554

Hubel, D. H., and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol*. 148, 574–591. doi: 10.1113/jphysiol.1959.sp006308

Ikegaya, Y., Aaron, G., Cossart, R., Aronov, D., Lampl, I., Ferster, D., et al. (2004). Synfire chains and cortical songs: temporal modules of cortical activity. *Science* 304, 559–564. doi: 10.1126/science.1093173

Izhikevich, E. M. (2007). *Dynamical Systems in Neuroscience*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/2526.001.0001

Jaeger, H. (2001). *The "Echo State" Approach to Analysing and Training Recurrent Neural Networks-With an Erratum Note*. Bonn: German National Research Center for Information Technology GMD Technical Report 148.

Ji, D., and Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci*. 10, 100–107. doi: 10.1038/nn1825

Jones, L. M., Fontanini, A., Sadacca, B. F., Miller, P., and Katz, D. B. (2007). Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proc. Natl. Acad. Sci. U.S.A*. 104, 18772–18777. doi: 10.1073/pnas.0705546104

Jouffroy, G. (2007). "Design of simple limit cycles with recurrent neural networks for oscillatory control," in *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)* (Cincinnati, OH: IEEE), 50–55. doi: 10.1109/ICMLA.2007.99

Kaplan, H. S., Thula, O. S., Khoss, N., and Zimmer, M. (2020). Nested neuronal dynamics orchestrate a behavioral hierarchy across timescales. *Neuron* 105, 562–576. doi: 10.1016/j.neuron.2019.10.037

Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A., and Arieli, A. (2003). Spontaneously emerging cortical representations of visual attributes. *Nature* 425:954. doi: 10.1038/nature02078

Kiebel, S. J., Daunizeau, J., and Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Comput. Biol*. 4:e1000209. doi: 10.1371/journal.pcbi.1000209

Kiebel, S. J., Von Kriegstein, K., Daunizeau, J., and Friston, K. J. (2009). Recognizing sequences of sequences. *PLoS Comput. Biol*. 5:e1000464. doi: 10.1371/journal.pcbi.1000464

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci*. 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Koechlin, E., Ody, C., and Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science* 302, 1181–1185. doi: 10.1126/science.1088545

Körding, K. P., and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature* 427:244. doi: 10.1038/nature 02169

Kotz, S. A., Meyer, M., Alter, K., Besson, M., von Cramon, D. Y., and Friederici, A. D. (2003). On the lateralization of emotional prosody: an event-related functional MR investigation. *Brain Lang.* 86, 366–376. doi: 10.1016/S0093-934X(02)00532-1

Koutnik, J., Greff, K., Gomez, F., and Schmidhuber, J. (2014). A clockwork RNN. *arXiv* 1402.3511.

Krause, J., Johnson, J., Krishna, R., and Fei-Fei, L. (2017). "A hierarchical approach for generating descriptive image paragraphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 317–325. doi: 10.1109/CVPR.2017.356

Kurata, G., Ramabhadran, B., Saon, G., and Sethy, A. (2017). Lan" guage modeling with highway LSTM," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (Okinawa: IEEE), 244–251. doi: 10.1109/ASRU.2017.8268942

Kurikawa, T., and Kaneko, K. (2015). Memories as bifurcations: realization by collective dynamics of spiking neurons under stochastic inputs. *Neural Netw.* 62, 25–31. doi: 10.1016/j.neunet.2014.07.005

Kurth-Nelson, Z., Economides, M., Dolan, R. J., and Dayan, P. (2016). Fast sequences of non-spatial state representations in humans. *Neuron* 91, 194–204. doi: 10.1016/j.neuron.2016.05.028

Laboy-Juárez, K. J., Langberg, T., Ahn, S., and Feldman, D. E. (2019). Elementary motion sequence detectors in whisker somatosensory cortex. *Nat. Neurosci.* 22, 1438–1449. doi: 10.1038/s41593-019-0448-6

Lashley, K. S. (1951). "The problem of serial order in behavior," in *Cerebral Mechanisms in Behavior; The Hixon Symposium*, ed L. A. Jeffress (Wiley), 112–146.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Leptourgos, P., Denève, S., and Jardri, R. (2017). Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? *Curr. Opin. Neurobiol.* 46, 154–161. doi: 10.1016/j.conb.2017.08.012

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* 31, 2906–2915. doi: 10.1523/JNEUROSCI.3684-10.2011

Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv* 1506.00019.

Litvak, V., Sompolinsky, H., Segev, I., and Abeles, M. (2003). On the transmission of rate code in long feedforward networks with excitatory-inhibitory balance. *J. Neurosci.* 23, 3006–3015. doi: 10.1523/JNEUROSCI.23-07-03006.2003

Liu, H., He, L., Bai, H., Dai, B., Bai, K., and Xu, Z. (2018). "Structured inference for recurrent hidden semi-Markov model," in *IJCAI* (Stockholm), 2447–2453. doi: 10.24963/ijcai.2018/339

Liu, P., Qiu, X., Chen, X., Wu, S., and Huang, X. (2015). "Multi-timescale long short-term memory neural network for modelling sentences and documents," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Okinawa), 2326–2335. doi: 10.18653/v1/D15-1280

Long, M. A., Jin, D. Z., and Fee, M. S. (2010). Support for a synaptic chain model of neuronal sequence generation. *Nature* 468:394. doi: 10.1038/nature09514

Lukoševičius, M., Jaeger, H., and Schrauwen, B. (2012). Reservoir computing trends. *Künstl. Intell.* 26, 365–371. doi: 10.1007/s13218-012-0204-5

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.* 14, 2531–2560. doi: 10.1162/089976602760407955

MacDonald, C. J., Lepage, K. Q., Eden, U. T., and Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron* 71, 737–749. doi: 10.1016/j.neuron.2011.07.012

Malhotra, P., Vig, L., Shroff, G., and Agarwal, P. (2015). "Long short term memory networks for anomaly detection in time series," in *Proceedings* (Louvain-la-Neuve: Presses Universitaires de Louvain), 89.

Martinez-Conde, S. (2006). Fixational eye movements in normal and pathological vision. *Prog. Brain Res.* 154, 151–176. doi: 10.1016/S0079-6123(06) 54008-7

Martinez-Conde, S., Macknik, S. L., and Hubel, D. H. (2004). The role of fixational eye movements in visual perception. *Nat. Rev. Neurosci.* 5, 229–240. doi: 10.1038/nrn1348

Mattar, M. G., Kahn, D. A., Thompson-Schill, S. L., and Aguirre, G. K. (2016). Varying timescales of stimulus integration unite neural adaptation and prototype formation. *Curr. Biol.* 26, 1669–1676. doi: 10.1016/j.cub.2016.04.065

Mazor, O., and Laurent, G. (2005). Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron* 48, 661–673. doi: 10.1016/j.neuron.2005.09.032

Meunier, D., Lambiotte, R., and Bullmore, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Front. Neurosci.* 4:200. doi: 10.3389/fnins.2010.00200

Meunier, D., Lambiotte, R., Fornito, A., Ersche, K., and Bullmore, E. T. (2009). Hierarchical modularity in human brain functional networks. *Front. Neuroinform.* 3:37. doi: 10.3389/neuro.11.037.2009

Mi, Y., Katkov, M., and Tsodyks, M. (2017). Synaptic correlates of working memory capacity. *Neuron* 93, 323–330. doi: 10.1016/j.neuron.2016.12.004

Miconi, T. (2017). Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *Elife* 6:e20899. doi: 10.7554/eLife.20899

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63:81. doi: 10.1037/h0043158

Mujika, A., Meier, F., and Steger, A. (2017). "Fast-slow recurrent neural networks," in *Advances in Neural Information Processing Systems*, 5915–5924.

Murray, J. D., Bernacchia, A., Freedman, D. J., Romo, R., Wallis, J. D., Cai, X., et al. (2014). A hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.* 17:1661. doi: 10.1038/nn.3862

Neverova, N., Wolf, C., Lacey, G., Fridman, L., Chandra, D., Barbello, B., et al. (2016). Learning human identity from motion patterns. *IEEE Access* 4, 1810–1820. doi: 10.1109/ACCESS.2016.2557846

Neves, F. S., and Timme, M. (2012). Computation by switching in complex networks of states. *Phys. Rev. Lett.* 109:018701. doi: 10.1103/PhysRevLett.109.018701

Nolfi, S. (2002). Evolving robots able to self-localize in the environment: the importance of viewing cognition as the result of processes occurring at different time-scales. *Connect. Sci.* 14, 231–244. doi: 10.1080/09540090208559329

O'Neill, J., Boccara, C., Stella, F., Schönenberger, P., and Csicsvari, J. (2017). Superficial layers of the medial entorhinal cortex replay independently of the hippocampus. *Science* 355, 184–188. doi: 10.1126/science.aag2787

Pastalkova, E., Itskov, V., Amarasingham, A., and Buzsáki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science* 321, 1322–1327. doi: 10.1126/science.1159775

Perdikis, D., Huys, R., and Jirsa, V. K. (2011). Time scale hierarchies in the functional organization of complex behaviors. *PLoS Comput. Biol.* 7:e1002198. doi: 10.1371/journal.pcbi.1002198

Pezzulo, G., van der Meer, M. A., Lansink, C. S., and Pennartz, C. M. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends Cogn. Sci.* 18, 647–657. doi: 10.1016/j.tics.2014.06.011

Pfeiffer, B. E. (2020). The content of hippocampal "replay". *Hippocampus* 30, 6–18. doi: 10.1002/hipo.22824

Prut, Y., Vaadia, E., Bergman, H., Haalman, I., Slovin, H., and Abeles, M. (1998). Spatiotemporal structure of cortical activity: properties and behavioral relevance. *J. Neurophysiol.* 79, 2857–2874. doi: 10.1152/jn.1998.79.6.2857

Rabinovich, M., Huerta, R., and Laurent, G. (2008). Transient dynamics for neural processing. *Science* 321, 48–50. doi: 10.1126/science.1155564

Rabinovich, M., Huerta, R., Volkovskii, A., Abarbanel, H., Stopfer, M., and Laurent, G. (2000). Dynamical coding of sensory information with competitive networks. *J. Physiol.* 94, 465–471. doi: 10.1016/S0928-4257(00)01092-5

Rabinovich, M., Volkovskii, A., Lecanda, P., Huerta, R., Abarbanel, H., and Laurent, G. (2001). Dynamical encoding by networks of competing neuron groups: winnerless competition. *Phys. Rev. Lett.* 87:068102. doi: 10.1103/PhysRevLett.87.068102

Rabinovich, M. I., Huerta, R., Varona, P., and Afraimovich, V. S. (2006). Generation and reshaping of sequences in neural systems. *Biol. Cybernet.* 95:519. doi: 10.1007/s00422-006-0121-5

Rahnev, D., and Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behav. Brain Sci.* 41, 1–107. doi: 10.1017/S0140525X18000936

Rajan, K., Harvey, C. D., and Tank, D. W. (2016). Recurrent network models of sequence generation and memory. *Neuron* 90, 128–142. doi: 10.1016/j.neuron.2016.02.009

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2:79. doi: 10.1038/4580

Ravasz, E., and Barabási, A. L. (2003). Hierarchical organization in complex networks. *Phys. Rev. E* 67:026112. doi: 10.1103/PhysRevE.67.026112

Reitich-Stolero, T., and Paz, R. (2019). Affective memory rehearsal with temporal sequences in amygdala neurons. *Nat. Neurosci.* 22, 2050–2059. doi: 10.1038/s41593-019-0542-9

Rivera, D. C., Bitzer, S., and Kiebel, S. J. (2015). Modelling odor decoding in the antennal lobe by combining sequential firing rate models with Bayesian inference. *PLoS Comput. Biol.* 11:e1004528. doi: 10.1371/journal.pcbi.1004528

Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., and Van Der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Hum. Mov. Sci.* 26, 525–554. doi: 10.1016/j.humov.2007.04.001

Schaal, S., Mohajerian, P., and Ijspeert, A. (2007). Dynamics systems vs. optimal control—a unifying view. *Prog. Brain Res.* 165, 425–445. doi: 10.1016/S0079-6123(06)65027-9

Schmidt, K. L., Ambadar, Z., Cohn, J. F., and Reed, L. I. (2006). Movement differences between deliberate and spontaneous facial expressions: zygomaticus major action in smiling. *J. Nonverb. Behav.* 30, 37–52. doi: 10.1007/s10919-005-0003-x

Seidemann, E., Meilijson, I., Abeles, M., Bergman, H., and Vaadia, E. (1996). Simultaneously recorded single units in the frontal cortex go through sequences of discrete and stable states in monkeys performing a delayed localization task. *J. Neurosci.* 16, 752–768. doi: 10.1523/JNEUROSCI.16-02-00752.1996

Sherman, S. M., and Guillery, R. (1998). On the actions that one nerve cell can have on another: distinguishing "drivers" from "modulators". *Proc. Natl. Acad. Sci. U.S.A.* 95, 7121–7126. doi: 10.1073/pnas.95.12.7121

Skaggs, W. E., and McNaughton, B. L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* 271, 1870–1873. doi: 10.1126/science.271.5257.1870

Soltani, A., Khorsand, P., Guo, C., Farashahi, S., and Liu, J. (2016). Neural substrates of cognitive biases during probabilistic inference. *Nat. Commun.* 7:11393. doi: 10.1038/ncomms11393

Soltani, R., and Jiang, H. (2016). Higher order recurrent neural networks. *arXiv* 1605.00064.

Stachenfeld, K. L., Botvinick, M. M., and Gershman, S. J. (2017). The hippocampus as a predictive map. *Nat. Neurosci.* 20:1643. doi: 10.1038/nn.4650

Strogatz, S. H. (2018). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Boca Raton, FL: CRC Press. doi: 10.1201/9780429492563

Sussillo, D., and Abbott, L. F. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron* 63, 544–557. doi: 10.1016/j.neuron.2009.07.018

Taherkhani, A., Belatreche, A., Li, Y., Cosma, G., Maguire, L. P., and McGinnity, T. M. (2020). A review of learning in biologically plausible spiking neural networks. *Neural Netw.* 122, 253–272. doi: 10.1016/j.neunet.2019.09.036

Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., et al. (2019). Recent advances in physical reservoir computing: a review. *Neural Netw.* 115, 100–123. doi: 10.1016/j.neunet.2019.03.005

Tokdar, S., Xi, P., Kelly, R. C., and Kass, R. E. (2010). Detection of bursts in extracellular spike trains using hidden semi-Markov point process models. *J. Comput. Neurosci.* 29, 203–212. doi: 10.1007/s10827-009-0182-2

Toutounji, H., and Pipa, G. (2014). Spatiotemporal computations of an excitable and plastic brain: neuronal plasticity leads to noise-robust and noise-constructive computations. *PLoS Comput. Biol.* 10:e1003512. doi: 10.1371/journal.pcbi.1003512

Tully, P. J., Lindén, H., Hennig, M. H., and Lansner, A. (2016). Spike-based Bayesian-Hebbian learning of temporal sequences. *PLoS Comput. Biol.* 12:e1004954. doi: 10.1371/journal.pcbi.1004954

Ulrych, T. J., Sacchi, M. D., and Woodbury, A. (2001). A bayes tour of inversion: a tutorial. *Geophysics* 66, 55–69. doi: 10.1190/1.1444923

VanRullen, R., and Koch, C. (2003). Is perception discrete or continuous? *Trends Cogn. Sci.* 7, 207–213. doi: 10.1016/S1364-6613(03)00095-0

Varona, P., Rabinovich, M. I., Selverston, A. I., and Arshavsky, Y. I. (2002). Winnerless competition between sensory neurons generates chaos: a possible mechanism for molluscan hunting behavior. *Chaos* 12, 672–677. doi: 10.1063/1.1498155

Watzenig, D. (2007). Bayesian inference for inverse problems-statistical inversion. *Elektrotech. Inform.* 124, 240–247. doi: 10.1007/s00502-007-0449-0

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nat. Neurosci.* 5, 598–604. doi: 10.1038/nn0602-858

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science* 269, 1880–1882. doi: 10.1126/science.7569931

Wörgötter, F., and Porr, B. (2005). Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. *Neural Comput.* 17, 245–319. doi: 10.1162/0899766053011555

Wu, Z., and King, S. (2016). "Investigating gated recurrent networks for speech synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 5140–5144. doi: 10.1109/ICASSP.2016.7472657

Yamashita, Y., and Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS Comput. Biol.* 4:e1000220. doi: 10.1371/journal.pcbi.1000220

Yan, S., Smith, J. S., Lu, W., and Zhang, B. (2018). Hierarchical multi-scale attention networks for action recognition. *Signal Process.* 61, 73–84. doi: 10.1016/j.image.2017.11.005

Yildiz, I. B., Jaeger, H., and Kiebel, S. J. (2012). Re-visiting the echo state property. *Neural Netw.* 35, 1–9. doi: 10.1016/j.neunet.2012.07.005

Yildiz, I. B., and Kiebel, S. J. (2011). A hierarchical neuronal model for generation and online recognition of birdsongs. *PLoS Comput. Biol.* 7:e1002303. doi: 10.1371/journal.pcbi.1002303

Yildiz, I. B., von Kriegstein, K., and Kiebel, S. J. (2013). From birdsong to human speech recognition: Bayesian inference on a hierarchy of nonlinear dynamical systems. *PLoS Comput. Biol.* 9:e1003219. doi: 10.1371/journal.pcbi.1003219

Yu, S. Z. (2015). *Hidden Semi-Markov Models: Theory, Algorithms and Applications*. Burlingotn, MA: Morgan Kaufmann.

Yu, Y., Si, X., Hu, C., and Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* 31, 1235–1270. doi: 10.1162/neco_a_01199

Zeki, S., and Shipp, S. (1988). The functional logic of cortical connections. *Nature* 335:311. doi: 10.1038/335311a0

Zemel, R. S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Comput.* 10, 403–430. doi: 10.1162/089976698300017818

Zen, H., Tokuda, K.,Masuko, T., Kobayashi, T., and Kitamura, T. (2004). "Hidden semi-markov model based speech synthesis," in *Eighth International Conference on Spoken Language Processing* (Jeju Island).

Zhang, C., and Woodland, P. C. (2018). "High order recurrent neural networks for acoustic modelling," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 5849–5853. doi: 10.1109/ICASSP.2018.8461608

Zheng, P., and Triesch, J. (2014). Robust development of synfire chains from multiple plasticity mechanisms. *Front. Comput. Neurosci.* 8:66. doi: 10.3389/fncom.2014.00066

Zutshi, I., Leutgeb, J. K., and Leutgeb, S. (2017). Theta sequences of grid cell populations can provide a movement-direction signal. *Curr. Opin. Behav. Sci.* 17, 147–154. doi: 10.1016/j.cobeha.2017.08.012

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership