A new predictive coding model for a more comprehensive account of delusions

Jessica Niamh Harding, Noham Wolpe, Stefan Peter Brugger, Victor Navarro, Christoph Teufel, Paul Charles Fletcher

Attempts to understand psychosis—the experience of profoundly altered perceptions and beliefs—raise questions about how the brain models the world. Standard predictive coding approaches suggest that it does so by minimising mismatches between incoming sensory evidence and predictions. By adjusting predictions, we converge iteratively on a best guess of the nature of the reality. Recent arguments have shown that a modified version of this framework—hybrid predictive coding—provides a better model of how healthy agents make inferences about external reality. We suggest that this more comprehensive model gives us a richer understanding of psychosis compared with standard predictive coding accounts. In this Personal View, we briefly describe the hybrid predictive coding model and show how it offers a more comprehensive account of the phenomenology of delusions, thereby providing a potentially powerful new framework for computational psychiatric approaches to psychosis. We also make suggestions for future work that could be important in formalising this novel perspective.

Lancet Psychiatry 2024

Published Online January 16, 2024 https://doi.org/10.1016/S2215-0366(23)00411-X

School of Clinical Medicine (J N Harding BA) **and Department of Psychiatry** (J N Harding, N Wolpe MD PhD, Prof P C Fletcher MBBS PhD), **University of Cambridge, Cambridge, UK; Department of Physical Therapy, The Stanley Steyer School of Health Professions, Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel** (N Wolpe)**; Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel** (N Wolpe)**; Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, Cardiff, UK** (S P Brugger MBBS MSc, V Navarro PhD, C Teufel PhD)**; Centre for Academic Mental Health, Bristol Medical school, University of Bristol, Bristol, UK** (S P Brugger)**; Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge, UK** (Prof P C Fletcher)**; Wellcome Trust MRC Institute of Metabolic Science, University of Cambridge, Cambridge, UK** (Prof P C Fletcher)

Correspondence to: Jessica Niamh Harding, School of Clinical Medicine, University of Cambridge, Cambridge CB2 0SZ, UK
jh2215@cam.ac.uk

Introduction

Computational psychiatry shares the aspirations of cognitive neuropsychology[1] and neuropsychiatry[2]: to scrutinise and understand symptoms of mental illness in the context of the functional attributes of the healthy systems that have been altered or perturbed.[2] Predictive coding models of psychiatric symptoms have had an important role in computational psychiatry, promising a principled description of how relatively simple disturbances in the brain's computations can account for an array of complex experiences.[3,4] These models offer a genuine hope for bridging the explanatory gap between neurobiological, cognitive, and subjective levels of explanation, and, ultimately, for a comprehensive multilevel perspective on the causes, emergence, and course of the symptoms of many psychiatric conditions.[5–8] This approach has been especially useful in previous attempts to understand psychosis: a complex experience of an alternative or unshared reality, characterised by delusions (ie, seemingly irrational but firmly held beliefs) and hallucinations (perceptions occurring in the absence of causative stimuli). By applying a model of healthy inference in the brain to explore psychosis, predictive coding models have been claimed to account for both delusions and hallucinations.[3]

The standard predictive coding framework suggests that the brain models reality through an iterative process, wherein top-down beliefs are adjusted to minimise prediction errors between incoming sensory evidence and the agent's predictions of this evidence. However, a modified version of this framework—hybrid predictive coding[9]—has been claimed to provide a better model of how healthy agents make inferences about external reality. This more comprehensive model could provide a greater understanding of the nature of psychosis, and allow researchers and clinicians to relate delusions and hallucinations to underlying neurobiology.

Explanations of delusions might benefit particularly from what the hybrid predictive coding model has to offer. The prodromal phase of psychosis and the initial emergence of delusional beliefs have been compellingly framed in terms of alterations in the iterative inference processes outlined by standard predictive coding models. However, several core aspects of established delusions, which are prominent in clinical practice, do not readily lend themselves to such an account. For example, a delusional belief could appear suddenly as a revelatory, self-evident, all-explaining, and unshakeable insight that seems impervious to contradictory evidence. The suddenness with which delusional beliefs can arise is hard to reconcile with the process of iteratively refined inferences suggested by predictive coding. The standard predictive coding model struggles to account for the transition from the unstable prodromal phase to the stable delusional state, and for the fixedness of established delusional beliefs. How do we explain the sometimes very definite point at which a person moves from a search for explanations to a strong conviction that they have found the over-arching truth? Current models of delusions would be more valuable for both research and clinical practice if they were better able to provide mechanistic explanations for such key phenomenological features. This improved capability would also offer increased potential for inspiring treatments.

The hybrid predictive coding framework expands standard predictive coding by adding an amortisation component that provides the top-down process with an initial best guess from which to start the iterative process. This initial best guess is a rapid, gist-based route to inference, learned through experience; as such, it offers speed and efficiency in stable environments, but could fail in more volatile or changeable circumstances. We note that the amortised versus inference dichotomy might be successfully applied to accounts of delusions that go beyond predictive coding, but which fall within the wider field of modelling cognition and behaviour as arising from (approximately) Bayesian inference.[10–12] However, in this Personal View we focus particularly on predictive coding owing to its substantial influence on theoretical accounts of delusions, as well as wider psychotic

phenomena. We begin by briefly reviewing existing predictive coding accounts of delusions, highlighting their explanatory limitations. We then discuss the hybrid account and show how the addition of amortised inference provides a richer perspective on delusions, one that more comprehensively addresses the phenomenological aspects that have so far been neglected.

Predictive coding and delusions

The hierarchical predictive coding model posits that because the brain cannot directly access the external world, it infers the most likely cause of an experienced sensory input by using a hierarchically organised set of predictions.[13–19] This inferential process can be formalised in Bayesian terms: a probabilistic prediction (ie, a prior belief) is combined with the sensory data (likelihood) to compute the cause with the greatest posterior probability. Prediction error signals resulting from this process indicate that the brain's current inference about the world is inaccurate and might need updating. By iteratively updating prior beliefs to minimise prediction errors, the brain more accurately perceives and models the external world. This updating occurs both quickly, as part of the perceptual process in each interaction between the agent and its environment, and more slowly, as an update of the agent's generative model that occurs over many instances of such interactions. Estimated precision of prediction errors at different levels determines the extent to which prior beliefs are updated; the more precise the sensory data are estimated to be relative to the prior, the more the sensory-level prediction error will drive updating of the prior.

When applied to delusions, the predictive coding framework appeals to an imbalance in the integration of priors and evidence. The core atypicality is formulated as an increase in the estimated precision of the sensory evidence relative to the precision of the prior, causing the brain's model of reality to become perturbed through relying unduly on sensory inputs (rather than on pre-existing knowledge of the world).[3,13,20–23] This predictive coding model has provided compelling explanations for several features of delusions, such as the prodromal period,[24] in which there is excessive uncertainty, a sense of altered salience, and a feeling that new beliefs (ie, hypotheses) must be generated to encompass unusual experiences. Experientially, the relatively reduced precision of prior beliefs (compared with that of incoming sensory data) would lead to a sense of uncertainty and change: a powerful feeling that one's existing world model was no longer valid and that something needed to be explained. Patterns and coincidences would become salient and everyday occurrences might feel loaded with meaning. Therefore, a relatively low-level perturbation in the process of Bayesian inference can be related to the delusional mood (or Wahnstimmung)[25] that characterises the phenomenology of the prodromal period of psychosis.

Through its explanation of the prodromal phase, the predictive coding approach effectively offers computational explanations for subjective experiences. New psychiatric research has aimed to unite contemporary neuroscientific approaches to delusions with the phenomenological perspective.[26,27] An Editorial in *The Lancet Psychiatry*[28] highlights the importance of a phenomenological approach in potentially bridging divisions between patients and professionals. It is in the attempt to bridge this explanatory gap—a quest for consilience[29]—that computational psychiatry offers great potential.[30–34]

However, despite real progress in explaining some of the phenomenology of psychosis in computational terms, several key experiential features of delusions remain that cannot be so neatly explained by the predictive coding model. One key example is the model's difficulty in readily describing how the stable and entrenched delusion—a new set of beliefs that satisfies the quest for understanding and resolves the many puzzles of the prodromal phase—emerges from the state of profound uncertainty, flexibility, and explanation-seeking that is characteristic of the delusional mood. A predictive coding process emphasises that contradictory evidence is the drive to updating. So how might the same model explain fixed, impervious beliefs? Some suggestions appeal to a distinction between high-level and low-level prior beliefs, with high-level priors gaining additional weighting (or precision) as an adaptive response to noisy evidence.[3,35] A related argument is that entrenchment is a consequence of a persistent failure of the model to make error-free predictions, despite repeated updates.[36,37] These failures of predictions lead to a reduced learning rate and a growing tendency to ignore new evidence.[38] These suggestions, however, fail to explain the crucial shift in which a hunger for new priors (in the prodromal phase) is satisfied by a new belief system that is rapidly established and is soon unassailable (as in the delusional state). Moreover, such added-on explanations could call for caution as they highlight the poorly specified nature of a framework that seems to accommodate all possibilities.[39]

In addition to its difficulty in accounting for the point of shift from the prodrome to the delusional state, and the fixed beliefs of the delusional state itself, the predictive coding account struggles to explain the so-called insight experiences reported in clinical practice. Jenson, writing about his own experiences of schizophrenia, reports that "believing that one is influenced by an alien force does not have the experiential quality of a reasoned conclusion. Rather [...] there is an idea revealing itself, an idea that has its own or is its own stimulus. It carries a sense of truth and certainty; it is what it is by nature of itself. This idea is automatous and expresses a higher truth that supersedes one's own thoughts or knowledge about the ordinary world".[40] The relatively sudden and fully formed presentation of some delusional beliefs, accompanied by this sense of indubitability and revelation, is difficult to reconcile with

 Published online January 16, 2024 https://doi.org/10.1016/S2215-0366(23)00411-X

the core idea of predictive coding: an iterative process of error detection and updating.

In short, the standard predictive coding framework has promise, but in the face of a more comprehensive phenomenology of delusions, its shortcomings become apparent. Most importantly, existing predictive coding models fail to adequately explain the transition from the prodromal to the delusional state, in which the quest to update is replaced by a fixed belief, and are unable to capture situations in which delusions rapidly emerge as sudden and unassailable revelations that seem to pervade the person's entire view of the world and its possibilities. In the next section, we describe an extension of the predictive coding framework and suggest ways in which it can solve these problems.

The hybrid predictive coding model

The hybrid predictive coding model proposes that, alongside the cycle of inference and updating central to standard predictive coding, a system of amortised inference learns a direct mapping between inputs and beliefs to streamline the process of inference.[9] Rather than starting the iterative process from a belief picked at random, the iterative process instead begins from a belief at which previous iterative processes arrived under similar conditions.

To illustrate, consider this problem: if $2x=20$, what is x? Assuming a basic understanding of algebra, you will probably have answered $x=10$. What would you conclude if we attempt to convince you that $x=0$? You would probably insist that we are wrong. However, our answer is perfectly rational. Why? Because if we substitute x for 0, we produce 20=20. Both the rapid arrival at the answer and the subsequent difficulty in understanding the logic behind the conclusion (until it is explained) illustrate amortised inference. The learning of stable contingencies between inputs and their (inferred) causes enables rapid employment of the previously taught inference and prohibits iterative analysis of the conclusion settled upon. Note how the assumption that the problem was an algebraic one was made without reflection. However, once this algebraic context has been established, it sets limits on the space of possibilities that one entertains in reflecting on the problem; the subsequent inference becomes inflexible and cannot be overridden easily. Amortised inference can therefore be compared with the acquisition of a mental habit.[41]

According to the hybrid model, amortised inference maps the input to an initial belief state—the optimal starting inference or best guess—that is then refined by the subsequent iterative process via standard prediction error minimisation. Any prediction error between the amortised prediction and the refined posterior (which indicates an inaccurate first guess) is used to adjust the parameters of the amortised function. Therefore, after each completed inference, the amortised component tries to learn a mapping from the input to the inference that the iterative component has settled on. The better this mapping is, the less refinement is needed by the iterative component during the next interaction between the agent and the environment. Once an accurate mapping has been learned, inference can be achieved nearly instantaneously after a single feedforward sweep, without iterative refinement, thereby minimising computational costs.

Although amortised inference is efficient and rapid in stable environments, when data are sparse or when the environment is unstable the inference is susceptible to inaccuracies.[42] Environments in which input-cause contingencies can be learned reliably provide the opportunity for rapid and computationally cheap inference, whereas more dynamic and unpredictable environments require a shift towards the iterative approach, which allows greater accuracy but has additional computational costs. The hybrid model is inherently sensitive to how much it needs to rely on each component: a suboptimal initial best guess offered by the amortised component will generate prediction errors, which will call for refinement of this initial guess by the iterative component, thereby upregulating iterative inference.[9]

Given that the iterative component uses prior beliefs to predict sensory inputs, it is inherently sensitive to one's understanding of the current state of the world. By contrast, the amortised component ignores the question of whether a specific state of the world is probable, and simply establishes how likely an inferred state is to have caused the current input. This notion partly resembles the distinction made by Teufel and Fletcher[17] between two forms of prediction; one relating to prior knowledge of local, context-specific regularities (ie, so-called expectations), and the other to knowledge of global, context-independent regularities (constraints). In some ways, these constraints are analogous to cached predictions used in amortised inference: they are applied ineluctably and inflexibly, allowing for rapid, efficient inference, but are prone to error, particularly in unstable environments. This perspective invites consideration of amortised inferences as working in a bottom-up way, in the sense that they are applied as a function of data rather than as higher level beliefs.

To summarise, the hybrid predictive coding model suggests that belief formation is optimised by a joint system of fast and slow inference: assumptive amortised inferences dominate when the agent can capitalise on stable regularities, with a corrective iterative system evaluating these inferences against the refined integration of prior beliefs and sensory evidence. Amortised inference could be considered akin to habitual coding insofar as it reflects a learned mapping, obviating the need to engage in computationally expensive but redundant iterative inference.[43] The initial inferences generated by this mapping are then subject to the iterative inferential processes characteristic of the standard predictive coding model, and the ultimate

output of this complementary inferential process creates a mapping that is used for future amortised inference.

A hybrid predictive coding model of delusions

We now explore how the hybrid model could allow a fuller account of the phenomenology of delusions than that provided by the standard predictive coding model. Hybrid predictive coding was conceptualised as a general model of inference, and delusions have not yet been formally explored within this model. Nevertheless, we can consider two components of amortised inference, which, if pathological, might contribute to the emergence of delusions: first, the selection and application of the appropriate mapping function for amortised inference and, second, the interaction between the amortised prediction and the iterative process. Our proposal for how hybrid predictive coding might contribute to our understanding of delusions remains preliminary and conceptual, and we speculate on possible directions of future development in the panel.

We will first consider how the choice of mapping function might contribute to delusion formation. The hybrid model posits that previously reliable environmental regularities drive rapid, system-wide predictions via learned, feedforward mapping from specific sensory inputs to a hierarchical series of inferences. Although not part of the current hybrid predictive coding model, this idea suggests that we might need to develop a library of possible amortised input–cause mappings based on past successful inferences.

The success of amortised inference depends on the selection of the correct mapping in the correct context. If we are watching a horror film and a drawer is opened to reveal a bread knife, the inferences we make will be markedly different from those accompanying a similar perceptual experience when watching a cooking programme or seeing a knife at the home of a friend. In the context of the horror film, we simply know that the bread knife will play a sinister part in future events. The amortised chain of inferences proceeds unrectified by iterative analysis because our prior beliefs about ominous symbolism in horror films encompass such a conclusion. In delusions, a perturbation in the system that selects the mapping from the library might enable the same conclusion in an incorrect context. The person might, automatically and unquestioningly, infer ominous consequences from seeing a knife in the home of a friend.

What might underpin this wrong choice of mapping? We have discussed how the pathological iterative process proposed by the standard predictive coding model could give rise to the prodromal sense of uncertainty and search for understanding. Sips describes, from his personal experience of psychosis, how this delusional mood "makes one literally question everything".[51] If tendency to question everything reflects a highly uncertain (imprecise) set of prior beliefs, the iterative process will follow a path towards beliefs that perfectly explain incoming sensory evidence, disregarding the beliefs built through previous experience. Weak priors serve a purpose in increasing the sensitivity to new contingencies, but, if heightened excessively, one's past experience becomes an unreliable benchmark, conferring an increased tendency to accept an initial best guess generated by the amortised component. For instance, an incorrectly selected amortised mapping that attaches an ominous conceptual inference to the perceptual

***Panel:* Future directions**

- Formal computational modelling focusing on learning and optimal selection of the amortised mapping, and on how the amortised and iterative parts of the system interact (eg, whether the amortised encoder could generate training data for the iterative generative model).
- How does amortised inference (perhaps in interaction with iterative inferences) take uncertainty into account? Generalised predictive coding models iteratively estimate precision parameters.[44] A generalised hybrid predictive coding model could amortise the estimation of the precision of sensory data, in addition to the predictions themselves (see Kingma and Welling's preprint on how a variational autoencoder learns a generative model of training data[45]).
- Might fundamental alterations in inbuilt constraints[15] affect amortised precision estimation, leading to altered estimations of environmental volatility? For example, could specific changes in neural systems (eg, noradrenergic changes) be identified that create an automatic bias in amortised precision or uncertainty estimation?
- Despite growing information on the neural circuitry underlying standard predictive coding,[4,46] little is currently known about the neural underpinnings of amortised inference, or about how the interaction between the two components is modulated.
- Observing the effects of various pharmacological perturbations on predictive coding circuity, with the addition of an amortised component, could contribute to our understanding of how psychotomimetic drugs might act.[43]
- As with other developments of standard predictive coding, whether and how newer models truly supersede existing approaches in their distinctiveness, validity, and scope should be established. A failure to address such questions has led to serious criticisms.[47,48] The focus of future work should be not only on what the hybrid model adds to predictive coding accounts, but also on its relationship to related but alternative accounts of psychosis.[49,50]
- Given that a more comprehensive model of healthy functioning could enrich attempts to account for specific perturbations, the addition of an amortised component should be explored in models of other psychiatric conditions, such as anxiety and depression.[6–10]