Review

Computational psychiatry: the brain as a phantastic organ

Karl J Friston, Klaas Enno Stephan, Read Montague, Raymond J Dolan

Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK
(Prof K J Friston FRS,
Prof K E Stephan PhD,
Prof R Montague PhD,
Prof R J Dolan FRS);
Translational Neuromodeling Unit, Institute of Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland (Prof K E Stephan);
and Computational Psychiatry Unit, Virginia Tech Carilion Research Institute, Roanoke, VA, USA (Prof R Montague)

Correspondence to:
Prof Karl J Friston, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London WC1N 3BG, UK
k.friston@ucl.ac.uk

In this Review, we discuss advances in computational neuroscience that relate to psychiatry. We review computational psychiatry in terms of the ambitions of investigators, emerging domains of application, and future work. Our focus is on theoretical formulations of brain function that put subjective beliefs and behaviour within formal (computational) frameworks—frameworks that can be grounded in neurophysiology down to the level of synaptic mechanisms. Understanding the principles that underlie the brain's functional architecture might be essential for an informed phenotyping of psychopathology in terms of its pathophysiological underpinnings. We focus on active (Bayesian) inference and predictive coding. Specifically, we show how basic principles of neuronal computation can be used to explain psychopathology, ranging from impoverished theory of mind in autism to abnormalities of smooth pursuit eye movements in schizophrenia.

Introduction

Computational psychiatry uses formal models of brain function to characterise the mechanisms of psychopathology, usually in a way that can be described in computational or mathematical terms. Computational psychiatry has arrived:[1–3] the first international computational psychiatry meeting was held in 2013, and 2014 saw the inception of the Max Planck Society-University College London initiative on computational psychiatry and ageing research—and the first UK computational psychiatry course. Several computational psychiatry units are emerging worldwide.

In this Review, we aim to provide tangible examples of computational psychiatry and to explain how it motivates mechanistic research in systems neuroscience; research that is being, or will soon be, translated into clinical neuroscience. First, we consider the properties a computational formulation must possess to be useful in psychiatry. We focus on active inference or predictive coding as an example (panel). We emphasise the importance of active inference by showing how it contextualises other formal treatments. We conclude with some examples of how theoretical principles can unify apparently disparate aspects of psychiatric disorders. We will consider functional and dissociative symptoms, soft neurological signs in schizophrenia, interoceptive inference and autism, dysconnection models of delusional (false) beliefs, and formal models of interpersonal exchange. We chose these examples to show the breadth of psychopathology that can be understood in terms of one pathology; namely, false inference that can be ascribed to neuromodulatory failures at the synaptic level. This Review is prospective, in that most of the examples we consider relate to the promise of the future—much of the work that substantiates the points we make has yet to be undertaken.

The phantastic organ

Many formal or computational schemes could characterise psychopathology, ranging from parallel distributed processing or neural network theory and dynamical systems theory, to reinforcement learning and game theory. However, these theoretical frameworks do not address the central problem encountered in psychiatry—ie, the production of false beliefs. The problems that concern psychiatrists are, almost universally, abnormal beliefs and their behavioural sequelae (eg, dysmorphophobia, paranoid ideation, organised delusional systems, hopelessness, poor self-worth, suicidal intent, obsessional thoughts, disorientation, false memories, and so on). This fact demands computational frameworks that deal with inference or beliefs and their neurophysiological realisation.

Within cognitive neuroscience, a new theory is emerging that helps us to understand false beliefs and how these arise from pathophysiology at the synaptic level. This perspective shifts away from the brain as a passive filter of sensations (or an elaborate stimulus–response link) towards a view of the brain as a statistical organ that generates hypotheses or fantasies that are tested against sensory evidence. In short, the brain is now considered a phantastic organ (from Greek *phantastikos*, the ability to create mental images). For many people, this perspective can be traced back to Hermann von Helmholtz and the notion of unconscious inference: that is, a pre-rational mechanism by which visual impressions are formed (eg, the seemingly automatic but erroneous belief that the sun rises and sets in the sky, as opposed to the truth that the Earth rotates around it).[4] In the past decade, the basic idea has been formalised and generalised to include deep or hierarchical Bayesian inference about the causes of sensations and how these inferences induce beliefs, movement, and behaviour.[5–9]

Predictive coding and the Bayesian brain

Modern versions of Helmholtz's ideas are now among the most popular explanations for message passing in the brain and are usually portrayed in the setting of the Bayesian brain hypothesis as predictive coding.[9–12] Predictive coding is not a normative or descriptive scheme, it is a process theory with a biologically plausible basis—there is now much circumstantial

anatomical and physiological evidence for predictive coding in the brain.[12–16] In this scheme, neuronal representations in higher levels of cortical hierarchies generate predictions of representations in lower levels. These top-down predictions are compared with representations at the lower level to form a prediction error (associated with the activity of superficial pyramidal cells). This mismatch signal is passed back up the hierarchy, to update higher representations (associated with the activity of deep pyramidal cells). For example, a high-level representation of a face would predict visual features encoded in early visual

---

Panel: Glossary of terms

Aberrant salience
The aberrant assignment of too much salience or precision to cues and events in the environment—usually associated with a hyperdopaminergic state in the brain.

Active inference
A corollary of the free energy principle that describes the process of inferring the causes of (sensory) data, which have to be actively chosen or sampled.

Bayes' rule
A rule for updating beliefs based on the fact that the probability of two events occurring together is the probability of the first event, assuming the second has occurred, times the probability of the second event occurring (and vice versa).

Bayesian brain
The notion that neuronal computations comply with Bayesian principles, such that neuronal representations are updated to produce posterior beliefs or expectations (about the causes of sensations) by combining prior expectations and sensory data according to Bayes' rule.

Dysconnection
The notion that psychopathology results from dysfunctional integration of distributed neuronal processes, usually at the synaptic level. This concept should be contrasted with disconnection, which refers to the anatomical disruption of long-range (extrinsic) connections at the level of white-matter tracts and axonal processes.

Embodied
Embodied is usually used in the context of embodied cognition (or mind) to emphasise that all aspects of cognition and perception are shaped by the body and physical (or physiological) exchanges with the external (or internal) environment.

Exteroceptive
Exteroceptive sensations are generated from outside the body—eg, sensory input from the retina and somatosensory organs. Exteroception includes hearing, vision, and touch.

Free energy principle
A description of biological processes as minimising (variational) free energy—a statistical measure of the surprise or improbability of sensory data, under a model of how those data were generated.

Force-matching illusion
The illusion that forces that are generated by an individual's own action are perceptually smaller in magnitude than when the same force is generated by someone else.

Interoceptive
Interoceptive sensations are generated from within the body—eg, sensory input from chemoreceptors, the gastrointestinal tract, and some cutaneous receptors.

Itinerancy
A dynamical behaviour in which a system wanders through its state space in an itinerant fashion, exploring different states.

Message passing
A concept from computer science in which messages are exchanged among processes. In the brain, the messages correspond to neuronal spiking in axonal connections. The (neuronal) messages invoke neuronal processes determined by synaptic connections and microcircuits.

Optimum decision theory
A normative theory of choice and behaviour under uncertainty, based on Bayesian principles and rationality. It is also known as the game theory or Bayesian decision theory.

Precision
The inverse dispersion or variance of a variable. A precise variable will take a value near its mean with high confidence or certainty.

Predictive coding
A scheme to infer the causes of data streams that was developed originally for signal processing and uses a model of how the data were generated. Predictive coding is now one of the most popular metaphors for message passing in the brain.

Proprioceptive
Proprioceptive sensations report the state and motion of the body. They include nociception (pain) and equilibrioception (balance).

Reinforcement learning
A branch of machine learning inspired by behavioural psychology that models behaviour as the maximisation of expected reward.

Self-organised criticality
The tendency of coupled dynamical systems to move towards regimes of instability that are characterised by critical slowing and phase transitions (similar to edge of chaos).

Sensory attenuation
The transient suspension of attention to the sensory consequences of an action—eg, the attenuation of visual motion (optical flow) sensations during rapid eye movements.
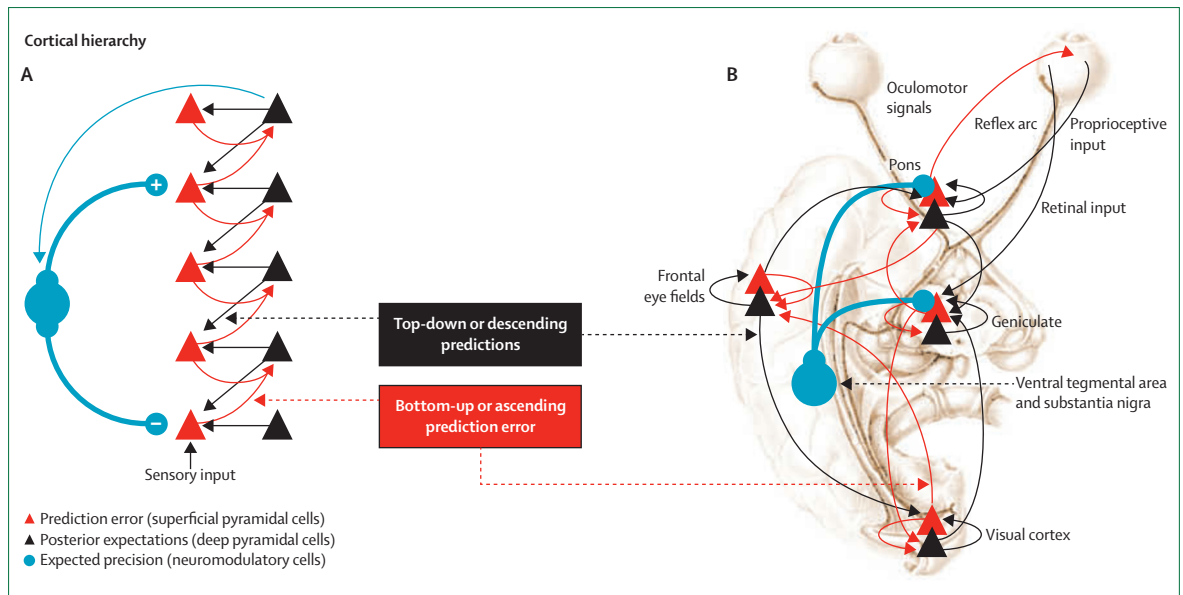
Figure 1: Hierarchical neuronal message passing system that underlies predictive coding
Neuronal activity encodes expectations about the causes of sensory input, and these expectations minimise prediction error. Minimisation relies on recurrent neuronal interactions between different levels of the cortical hierarchy. Within this model, the available evidence suggests that superficial pyramidal cells (red triangles) compare expectations (at each level) with top-down predictions from deep pyramidal cells (black triangles) at higher levels. (A) A simple cortical hierarchy with ascending prediction errors and descending predictions. Neuromodulatory gating or gain control (blue) of superficial pyramidal cells determines their relative influence on deep pyramidal cells encoding expectations. (B) Schematic example that shows the visual system. Putative cells of origin of ascending or forward connections convey prediction errors (red arrows) and descending or backward connections (black arrows) construct predictions. The prediction errors are weighted by their expected precision, which is associated with the activity of neuromodulatory systems—here, projections from ventral tegmental area and substantia nigra. In this example, the frontal eye fields send predictions to the primary visual cortex. However, the frontal eye fields also send proprioceptive predictions to pontine nuclei, which are passed to the oculomotor system to cause movement through classic reflexes. Here descending predictions to the visual cortex constitute corollary discharge. Every top-down prediction is reciprocated with a bottom-up prediction error to ensure predictions are constrained by sensory information. The resolution of proprioceptive prediction error is particularly important because it enables descending predictions (about the state of the body) to cause movement by dynamically resetting the equilibrium or set point of classic reflexes.

cortex. Any mismatch between descending predictions and low-level features will generate ascending prediction errors that might change the high-level representation into a smiling face. This recursive exchange of signals suppresses prediction error at each level of the hierarchy to provide a hierarchical explanation for sensory inputs at the lowest (sensory) level. In computational terms, neuronal activity is thought to encode beliefs or probability distributions over external states that cause sensations (eg, my visual sensations are caused by a face that is smiling). The simplest encoding associates the belief with the expected value of a (hidden) cause or expectation. These causes are referred to as hidden because they have to be inferred from their sensory consequences. For example, a smiling face is the hidden cause of visual sensations that has to be inferred from, and only from, visual input.

In summary, predictive coding represents a biologically plausible scheme that enables the brain to update beliefs about the world with sensory samples (figure 1). Neuroanatomy and neurophysiology can therefore be regarded as a distillation of causal structure that embodies a generative model, which produces predictions of sensations in a Helmholtzian sense. The implications for

Figure 2: Giuseppe Arcimboldo, *The Vegetable Gardener* (1590)
Arcimboldo used fruits and vegetables to create faces in his paintings.[17,18] Faces are probably one of the most important (hidden) causes of sensations.[19]

perception are intimated nicely in Arcimboldo's vegetable garden (figure 2) and in Kandel's[17] discussion of the beholder's share:

"The insight that the beholder's perception involves a top-down inference convinced Gombrich that there is no 'innocent eye'—that is, all visual perception is based on classifying concepts and interpreting visual information. One cannot perceive that which one cannot classify."[17]

From the Bayesian brain to active inference

If the brain is a generative model of the world, then much of it must be occupied by modelling other people. In other words, individuals spend most of their time predicting the internal (proprioceptive) and external (exteroceptive) consequences of behaviour (both their own and that of others). To fully appreciate the bilateral nature of these predictions, inference can be considered in an embodied context. In this setting, perception can be understood as resolving exteroceptive prediction errors by selecting the predictions that best explain sensations. Conversely, behaviour suppresses proprioceptive prediction error by changing proprioceptive sensations. This suppression relies on classic reflexes, in which equilibrium points are set by descending proprioceptive predictions (figure 1).[15] This process is called active inference, which involves equipping a predictive coding scheme with reflexes (figure 1).

If high-level sensorimotor expectations provide top-down predictions of the sensory consequences of moving, then the brain implicitly provides a range of hypotheses to infer the intentions of others. The exteroceptive (eg, visual) outcomes of another person's movements can be predicted from the results of making the same purposeful movement oneself—all that has to be done is to infer who is moving (self *vs* other). This notion might be important for understanding false beliefs about agency in schizophrenia[20] and provides an account of mirror neurons that respond to self-made acts and during action observation.[21] However, to harness the mirror neuron system during action observation, proprioceptive prediction errors have to be attenuated because they would otherwise elicit movements in the observer that mirror the subject of observation (as in echopraxia). This attenuation rests on reducing the influence of proprioceptive prediction errors, where this influence is determined by their precision. Precision can be regarded as a measure of signal-to-noise, or the confidence assigned to an information stream. Estimating precision is a fundamental aspect of inference in the brain, and can be regarded as encoding the expected uncertainty in any given context.[22–24] This estimation represents a subtle but generic problem that the brain must solve and the solution might rest on modulating the gain or excitability of neuronal populations that generate prediction error.[12,25,26]

Ascending prediction errors in cortical hierarchies can be regarded as broadcasting newsworthy information that has yet to be explained by descending predictions. However, the brain also has to select the channels it listens to. It can do this by adjusting the volume or gain of prediction errors that update expectations. This precision weighting of prediction errors is thought to be a generic computational process operating throughout the brain[24] and might be mediated by neuromodulatory mechanisms of gain control at a synaptic level.[27] In short, neuromodulatory gain control corresponds to a (Bayes-optimal) encoding of precision by the excitability of neuronal populations that report prediction errors. This cast computational light on why superficial pyramidal cells have many synaptic gain-control mechanisms such as those involving NMDA receptors and classic neuromodulatory receptors such as D1 dopamine receptors.[28–31] Furthermore, predictive coding places cortical excitation–inhibition balance in a key position to mediate precision engineered message passing within and among hierarchical levels.[32] This aspect of predictive coding has been associated with attentional gain control in sensory processing[25,33] and has been discussed in terms of affordance in active inference and action selection.[34–36] Crucially, the delicate balance of precision at different hierarchical levels has a profound effect on inference and could hold the key for a formal understanding of false beliefs in psychopathology.[37]

Interoceptive inference

Recently, investigators described emotional processing in terms of predictive coding or inference about interoceptive or bodily states.[38,39] In active inference, motor reflexes are driven by proprioceptive prediction errors. Proprioceptive prediction errors compare primary afferents from stretch receptors with proprioceptive predictions that descend to α motor neurons in the spinal-cord and cranial nerve nuclei. This circuit effectively replaces descending motor commands with proprioceptive predictions, which are fulfilled by peripheral reflexes.[15] Descending predictions rely on deep hierarchical inference about states of the world, including an individual's own body. Crucially, a similar principle can be invoked to explain homoeostasis (the control of blood pressure, glycaemia, etc), in which descending interoceptive predictions control autonomic reflexes. As with the mirror neuron system, interoceptive predictions constitute just one stream of multimodal predictions that are generated by expectations about the embodied self. The extension of active inference to include autonomic reflexes and interoceptive predictions raises many interesting questions.[19] For example, what role do neuromodulators such as dopamine and oxytocin have in mediating the precision of prediction errors? What is the relationship between exteroception and interoception during self-observation?[40] Do von Economo neurons convey interoceptive predictions from the insular cortex to the amygdala?[41] Researchers in neuropsychoanalysis are also asking key questions about hierarchical inference and emotional regulation.[42]

The Bayesian brain and other formal theories
Parallel-distributed processing, precision, and the dysconnection hypothesis

Part of the construct validity of active inference is that it leads to, and contextualises other formal approaches.

For example, formal models of schizophrenia are often described in terms of neuronal disconnection. There are two versions of the disconnection hypothesis: the first is implied by Wernicke's sejunction hypothesis, which postulates an anatomical disruption or disconnection of association fibres;[43] the second postulates abnormalities at the level of synaptic efficacy and plasticity, leading to dysfunctional integration or dysconnectivity among cortical and subcortical systems.[44] Dysfunctional integration at the synaptic level coincides with aberrant neuromodulatory precision control in predictive coding, which relates closely to theories framed in terms of signal-to-noise.[29,45,46] Furthermore, this putative abnormality is consistent with nearly every synaptic or physiological theory of schizophrenia, including a dysfunction of dopaminergic and NMDA receptors,[47,48] GABAergic abnormalities,[49,50] and dysfunctional excitation–inhibition balance.[51] A common theme in these theories is failure to maintain the appropriate gain of principal or pyramidal cells.

Reinforcement learning, game theory, and metacognition

Other examples of formal theories in psychiatry include reinforcement learning and optimum decision theory. These frameworks have shaped many aspects of systems neuroscience over the past decade.[52,53] They provide a normative account of choice behaviour with clear neurobiological correlates.[54] Perhaps the most celebrated correlate is the association between dopamine and reward prediction error in temporal difference models of reward learning,[52,55] However, there are many studies suggesting that dopamine also encodes precision or uncertainty.[56–61] In active inference, reward and value are treated as prior beliefs that determine predictions, including behaviour—eg, the taste of wine if I raise a glass to my lips. Prior expectations therefore represent the reward (in reinforcement learning) or utility (in behavioural economics). In this setting, a negative reward-prediction error signals a loss of confidence or precision in expectations about rewarding outcomes.[62] Applications of reinforcement learning to psychopathology have been framed in terms of Bayesian inference and prior beliefs.[63] Perhaps more tellingly, formal theories of aberrant salience (originally based on notions from reinforcement learning[64]) are now more commonly framed as theories of aberrant precision; particularly given the formal connection between precision control and attention or salience (panel).[25] The focus on aberrant precision also speaks to theories of metacognition, which are prevalent in psychology and social neuroscience. Metacognition, or the study of beliefs about beliefs, often focuses on reporting the confidence in decisions.[65] Metacognition affords a measure of a person's insight that has clear relevance for psychiatry and a direct link to subjective certainty or precision.[66]

A subtle but important advantage of describing reward or value functions in terms of prior beliefs (and the precision that is afforded those beliefs) is that the beliefs that account for individual behaviour (choice) can be defined, enabling quantitative and formal phenotyping in terms of beliefs and attitudes.[53,67] This approach forms the basis of many current computational psychiatry initiatives,[68] and has even been extended to game theory models of interpersonal exchange. These extensions might be important, particularly to characterise various psychopathies and their genetic or physiological correlates.[68,69]

Dynamical systems theory and self-organisation

Finally, there has been growing interest in characterising brain dynamics using concepts from dynamical systems theory such as criticality (panel).[70–72] Indeed, changes in the nature and deployment of coherent or coordinated dynamics have often been associated with disorders such as schizophrenia as evidence of functional dysconnection.[74] Criticality and itinerant dynamics relate to inference by providing a rich dynamical repertoire that enables the brain to respond quickly to changing inputs.[75] This dynamical repertoire provides a wide range of hypotheses that can be used to explain sensory data.[18,76]

If precision is estimated by the brain, and precision has a profound effect on dynamical stability,[77] then self-organised criticality is a necessary aspect of the functioning brain. It is this delicate control of instability that makes precision a key parameter in neuronal dynamics.

In the next section, we look at what would happen if the estimation of precision was compromised. We focus not on normative models (that describe what the brain does), but on process models (that describe how the brain does it), because our aim is to link the phenomenology of psychiatric disorders to their neurophysiological and molecular causes.[78]

Neuromodulation and false inference

We have introduced a computational framework for action and perception, with a special focus on the synaptic mechanisms that might underlie false inference in psychiatric disorders: in brief, the formal constraints implicit in predictive coding mandate modulatory gain control for ascending prediction errors. In an article in 2012, Edwards and colleagues[79] illustrates how functional symptoms can be understood as false inference about the causes of abnormal sensations, movements, or their absence. This example offered a simple neurophysiological explanation of symptoms that would otherwise be difficult to diagnose or formulate. This theme is repeatedly emerging in psychiatry, from false inference as an account of positive symptoms (hallucinations and delusions) in schizophrenia,[80] to the loss of central coherence in autism.[81] Moreover, it is remarkable that the

same role for precision weighting of prediction errors emerges from different theoretical treatments of learning and inference in the brain, including predictive coding in vision,[11] free-energy accounts of perception and behaviour,[7] and hierarchical Bayesian models of learning.[82]

Aberrant precision and sensory attenuation

A recurrent theme in many psychiatric disorders is a failure of sensory attenuation, with secondary consequences for the acquisition and deployment of hierarchically deep models of the world, and interpersonal interactions. In the context of sensory exchanges with the world, such as pursuit eye movements, a failure of sensory attenuation means that sensory precision is too high in relation to the precision of higher (prior) beliefs about the causes of sensations. It is relatively easy to reproduce the key deficits of slow pursuit eye movements in schizophrenia by simply reducing prior precision in simulations of eye tracking using predictive coding and oculomotor reflexes.[83] This mechanism might explain the inability of patients with schizophrenia to infer regular (high-order) contingencies that underlie target movement and anticipate its motion (this failure can be revealed with use of a mask or occluder, so that the target's reappearance from behind the occluder has to be anticipated). Because prior expectations are compromised in schizophrenia, violations (eg, unpredicted changes in target motion) paradoxically improve pursuit performance, relative to people without schizophrenia. This is because prior expectations about target motion are violated and confound the tracking behaviour of people with normal priors.[83] This example makes the more general point that the relative precision of sensory and prior prediction errors is a crucial determinant of a person's susceptibility to illusions and their responses to unpredicted events or their omissions.[37] Relatively simple, well established frameworks such as slow pursuit eye movements, the mismatch negativity, and the psychophysics of illusions might therefore be particularly useful in psychiatric phenotyping, because behavioural and neuronal responses can be characterised in terms of precision in hierarchical predictive coding.[24,27,84]

Crucially, quantities such as prediction error and precision have clear neurobiological correlates that allow modelling. For example, if prediction errors are reported by superficial pyramidal cells in the cortex, then prediction errors can be measured by non-invasive electromagnetic techniques because these cells contribute most to event-related potentials and induced responses.[85] Similarly, if precision is encoded by the excitability or gain of superficial pyramidal cells, then this gain can be estimated with use of biophysical modelling of neuronal circuits (dynamic causal modelling) on the basis of evoked electrophysiological responses.[86] Much work is being done on paradigms

that elicit prediction errors to characterise hierarchical message passing in people with psychiatric disorders.[87] Studies such as these could provide a computationally and biophysically grounded phenotype of psychiatric dysconnection syndromes.[84] These paradigms could then be used with psychopharmacological manipulations and genetic studies to identified the precise synaptic mechanisms for disease and their molecular basis. Although these (low-level) models of eye movements or neuronal responses to violations exploit the formal constraints offered by computational psychiatry, they do not touch on the deeper (high-level) beliefs that characterise psychosis.

Synthetic delusions

Simulation of delusional beliefs is straightforward because hierarchical Bayesian inference schemes such as predictive coding deal explicitly with expectations. Perhaps the best example addresses beliefs about agency—a key issue in schizophrenia research. Some patients with psychiatric disorders fail to contextualise the consequences of their actions and make false inferences about the agency or authors of their sensory outcomes. This is demonstrated nicely by the resistance of patients with schizophrenia to the force-matching illusion.[88,89] Normally, people show sensory attenuation when they do something, whereas patients with schizophrenia seem not to. The force-matching illusion reduces the perceived magnitude of self-produced forces relative to externally generated forces. Crucially, patients with schizophrenia are resistant to this illusion and can accurately report the forces that they produce themselves.[88] This result can be simulated in predictive coding of somatosensory and proprioceptive cues by, precluding an attenuation of sensory precision.[23] However, this comes at a price—to produce the self-generated force in the first place, non-sensory (or prior) precision must be increased so that an individual's prior belief that they are moving over-rides the sensory evidence that they are not. The problem here is that to explain the precise sensory information (that the force is always less than predicted) the person has to infer an opposing external force. This scenario is a good example of a simulated delusional belief that rests on one simple manipulation, a failure to attenuate sensory precision and compensatory increases in precision at higher levels of the hierarchy (figure 3).

Researchers are now starting to report abnormalities in the gain control of pyramidal cell populations in cortical hierarchies. For example, Fogelson and colleagues[90] used event-related potentials and dynamic causal modelling to show "the differences between recurrent inhibitory connections during the processing of predictable and unpredictable stimuli were markedly attenuated" in people with schizophrenia. Similarly, dynamic causal modelling of functional MRI signals suggests a selective reduction in recurrent inhibitory

connections within the medial prefrontal cortex, which again speaks to a failure to maintain the cortical gain control implicit in the aberrant encoding of precision.[91] Biophysical models—used to characterise evoked responses (and spontaneous activity)—could become increasingly defined in functional (computational) terms.

Autism and interoceptive inference

Perhaps the best example of computational approaches to neurodevelopmental syndromes comes from research on autism. Much of the phenomenology of autism has been described in terms of false (Bayesian) inference that results from a loss of prior precision, relative to sensory precision.[81,92,93] In autism, consequences of increases in (or a failure to attenuate) sensory precision are being interpreted in a developmental context. This interpretation is particularly interesting in light of interoceptive inference because it addresses the acquisition of generative models that distinguish between self and other. It could be that a failure to contextualise interoceptive cues, elicited by maternal interactions, precludes a proper attribution of agency to the interoceptive consequences of affiliative interactions; in other words, the infant with autism cannot learn to distinguish between autonomic responses elicited by the mother and those caused by its own interoceptive predictions. This concept has several interesting implications for attachment, theory of mind, and the poor central coherence that affects people with autism in adulthood.[94] It also provides an interesting explanation for interoceptive hypersensitivity (an autonomic over-responsiveness to interoceptive cues) in autism and a failure to engage with prosocial (exteroceptive) cues.[95] If this explanation is correct, then it provides a clear pointer to abnormalities of (precision) gain control in cortical systems that mediate interoceptive inference (such as the anterior insular and cingulate cortex[96,97]) and transactions with others.[53] Formal theories of autism highlight the importance of inferences about
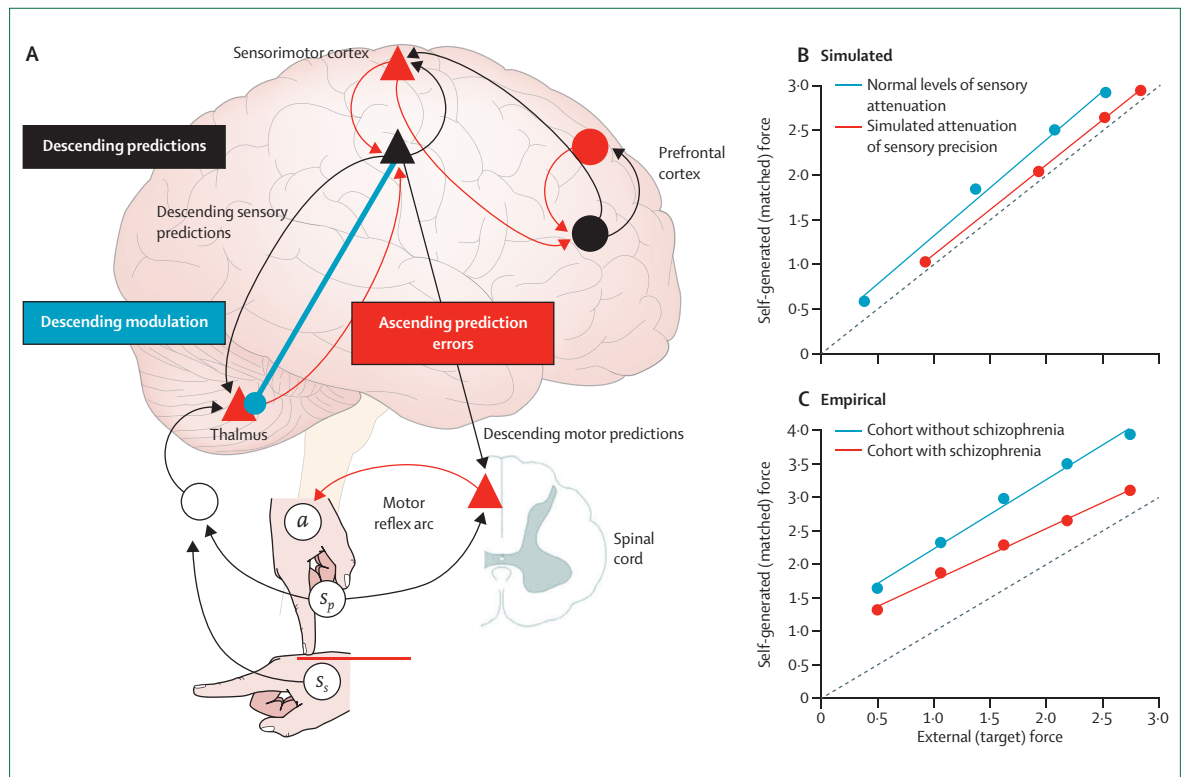
Figure 3: Predictive coding model in the force-matching illusion

(A) Shows a schematic of the predictive coding model used to simulate delusions and failure of the force-matching illusion in terms of aberrant precision. Somatosensory and proprioceptive prediction errors are generated by the thalamus, whereas the expectations and prediction errors about hidden causes (forces) are in sensorimotor and prefrontal cortex. Under active inference, proprioceptive predictions descend to the spinal cord and elicit output from alpha motor neurons (proprioceptive prediction-error units) via a classic reflex arc. As in figure 1, red connections mediate ascending prediction errors and black connections mediate descending predictions. The blue connection denotes descending neuromodulatory (eg, NMDA receptor) effects that mediate sensory attenuation. (B) The results of a force-matching simulation that was repeated under different levels of self-generated force induced by prior beliefs about hidden causes. For normal levels of sensory attenuation, the internally matched force was higher than was the externally generated force. Data from patients with schizophrenia were simulated by attenuating sensory precision and increasing the precision of prediction errors at higher levels of the hierarchy. This resulted in a more accurate perception of internally generated force (red). (C) Equivalent data from the force-matching task from controls and a cohort of patients with schizophrenia. Adapted from references 37.

others, a potentially important facet of computational psychiatry in its own right.

Interpersonal inference

Optimum Bayesian decision theory (or game theory) provides a potentially important framework to quantify beliefs about other people that affect interpersonal exchange. Indeed, game theory has already proved useful in the characterisation of autism.[67] Game theory is important because it allows the characterisation of interpersonal behaviour in terms of prior beliefs or expectations[82] and for researchers to ask whether these differ systematically among different personality traits or psychiatric diagnoses. One can use several relatively simple models—eg, the rock-paper-scissors and beads tasks, which reveal a tendency for people with schizophrenia to reach conclusions prematurely.[98–100] The putative role of dopamine in encoding the precision of beliefs about desired outcomes is of particular interest.[62] This line of thinking suggests that (economic) games can be used to formally characterise behaviour in terms of prior beliefs about outcomes and confidence in those beliefs.[69]

We chose the above examples to show how a formal approach can provide generic explanations for psychopathology that are physiologically grounded. However, one might ask if the notion of aberrant precision (ie, neuromodulation) is so inclusive that it is non-specific? Clearly, to understand the specific ways that aberrant precision is expressed, one has to understand the myriad of neuromodulatory mechanisms (eg, different neurotransmitters acting on different receptor subtypes in different parts of the brain) in relation to functional anatomy and neurodevelopment. This understanding might be necessary to understand the diversity of psychiatric disorders and the mechanistic basis of their classification.

Conclusion

In this Review, we have discussed how computational psychiatry can use formal models of perceptual inference and learning to provide a mechanistic and functional perspective on psychopathology and its underlying pathophysiology. We focused on inference as the overarching theoretical framework; largely
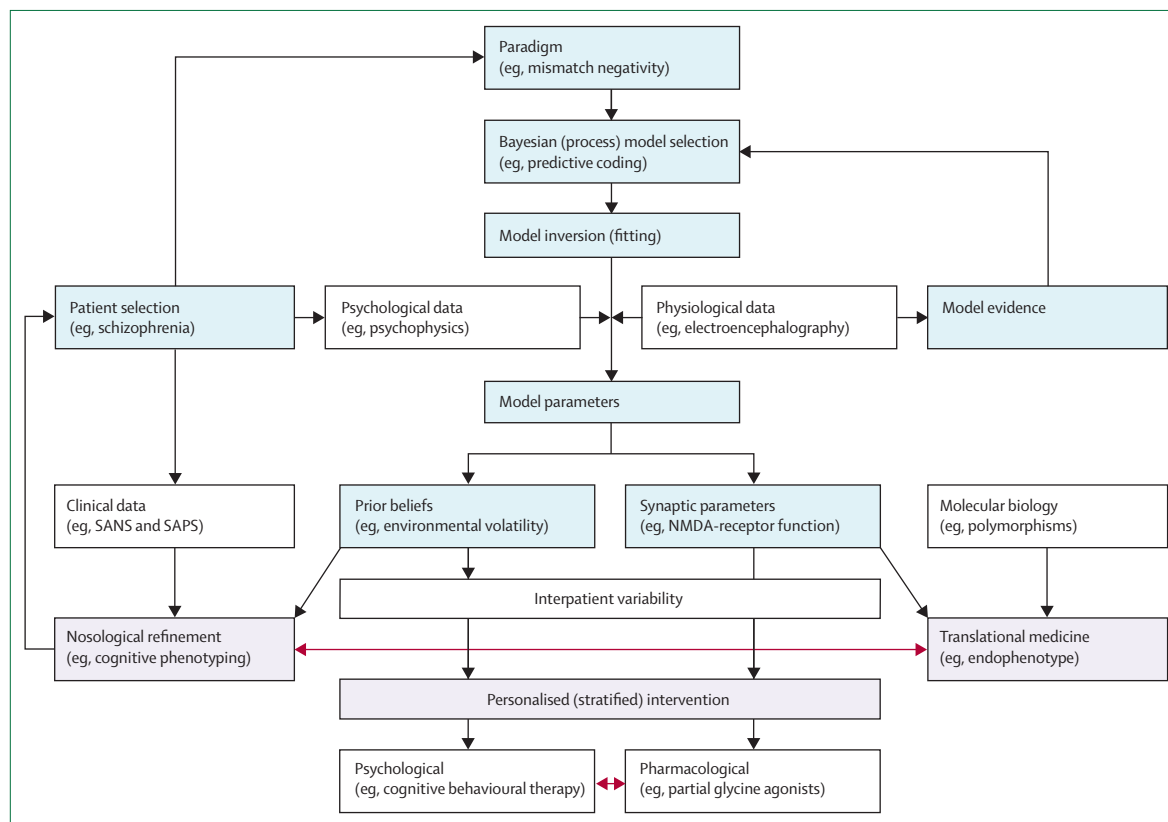
***Figure 4:* Formal (computational) models for nosological (or normative) descriptions of psychopathology**
Much of this roadmap is common sense (and speculative) but we try to make the point that formal models with a process theory (eg, predictive coding) can make bilateral predictions about behavioural and neuronal responses (eg, mismatch negativity responses to oddball stimuli). As such, they can be used as observation models of empirical (psychophysical and physiological) data. This (model inversion) furnishes model evidence—for selecting among competing models or hypotheses—and model parameters that quantify an individual's beliefs and their neuronal encoding (eg, NMDA receptor dependent plasticity in the oddball paradigm). In turn, these can be used to exploit individual variability for stratified psychological or pharmacological therapy. SANS=Scale for the Assessment of Negative Symptoms. SAPS=Scale for the Assessment of Positive Symptoms.

because it can formalise perception and behaviour in terms of probabilistic beliefs. By assuming that the brain engages in some form of active inference, neuronal dynamics and message passing can be associated with Bayesian belief updating. This enables some remarkably specific predictions about the effect of functional or synaptic dysconnections. We focused on neuromodulatory failures and how they can be understood in terms of an aberrant encoding of subjective precision or uncertainty, leading to false inference that can be expressed at many different levels. This computational approach necessarily enforces a mechanistic and quantitative view of psychopathology—a view that can accommodate phenomenology ranging from soft neurological signs in schizophrenia to theory of mind in autism, using exactly the same computational principles. The quantitative (and parametric) characterisations offered by computational psychiatry could enrich traditional psychiatric classification. Furthermore, the use of formal models might lead to levels of description that might, or might not, be appropriate for particular disorders. The roadmap for computational psychiatry (figure 4) highlights the integrative role of formal (process) models in establishing construct validity among nosological constructs and underlying neuronal processes and shows how this link might translate into therapy. Although roadmaps like this can be compelling, their navigation is usually a slow and challenging process.

Contributors
All authors contributed equally to the ideas and structure of this review. KJF took primary responsibility for writing. The text was subsequently revised by all authors.

Declaration of interests
We declare no competing interests.

References
1 Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. *Trends Cogn Sci* 2012; **16:** 72–80.
2 Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. *Nat Neurosci* 2011; **14:** 154–62.
3 Huys Q J, Moutoussis M, Williams J. Are computational models of any use to psychiatry? *Neural Netw* 2011; **24:** 544–51.
4 Helmholtz H. Concerning the perceptions in general. In: Treatise on physiological optics, 3rd edn. New York: Dover, 1866.
5 Dayan P, Hinton GE, Neal RM, Zemel RS. The Helmholtz machine. *Neural Comput* 1995; **7:** 889–904.
6 Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis* 2003; **20:** 1434–48.
7 Friston K, Kilner J, Harrison L. A free energy principle for the brain. *J Physiol Paris* 2006; **100:** 70–87.
8 Hohwy J. The Predictive Mind. Oxford: Oxford University Press, 2013.
9 Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci* 2013; **36:** 181–204.
10 Srinivasan MV, Laughlin SB, Dubs A. Predictive coding: a fresh view of inhibition in the retina. *Proc R Soc Lond B Biol Sci* 1982; **216:** 427–59.
11 Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* 1999; **2:** 79–87.
12 Friston K. Hierarchical models in the brain. *PLoS Comput Biol* 2008; **4:** e1000211.
13 Mumford D. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol Cybern* 1992; **66:** 241–51.
14 Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. *Neuron* 2012; **76:** 695–711.
15 Adams RA, Shipp S, Friston KJ. Predictions not commands: active inference in the motor system. *Brain Struct Funct* 2013; **218:** 611–43.
16 Shipp S, Adams RA, Friston KJ. Reflections on agranular architecture: predictive coding in the motor cortex. *Trends Neurosci* 2013; **36:** 706–16.
17 Kandel E. The Age of insight: the quest to understand the unconscious in art, mind, and brain, from Vienna 1900 to the present. New York: The Random House Publishing Group, 2012.
18 Gregory RL. Perceptual illusions and brain models. *Proc R Soc Lond B Biol Sci* 1968; **171:** 279–96.
19 Friston KJ. The fantastic organ. *Brain* 2013; **136:** 1328–32.
20 Frith CD, Blakemore SJ, Wolpert DM. Abnormalities in the awareness and control of action. *Philos Trans R Soc Lond B Biol Sci* 2000; **355:** 1771–88.
21 Kilner JM, Friston KJ, Frith CD. Predictive coding: an account of the mirror neuron system. *Cogn Process* 2007; **8:** 159–66.
22 Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. *Neuron* 2005; **46:** 681–92.
23 Brown H, Adams RA, Parees I, Edwards M, Friston K. Active inference, sensory attenuation and illusions. *Cogn Process* 2013; **14:** 411–27.
24 Iglesias S, Mathys C, Brodersen KH, et al. Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron* 2013; **80:** 519–30.
25 Feldman H, Friston KJ. Attention, uncertainty, and free-energy. *Front Hum Neurosci* 2010; **4:** 215.
26 Clark A. The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Front Psychol* 2013; **4:** 270.
27 Moran RJ, Campo P, Symmonds M, Stephan KE, Dolan RJ, Friston KJ. Free energy, precision and learning: the role of cholinergic neuromodulation. *J Neurosci* 2013; **33:** 8227–36.
28 Goldman-Rakic PS, Lidow MS, Smiley JF, Williams MS. The anatomy of dopamine in monkey and human prefrontal cortex. *J Neural Transm Suppl* 1992; **36** (suppl)**:** 163–77.
29 Braver TS, Barch DM, Cohen JD. Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. *Biol Psychiatry* 1999; **46:** 312–28.
30 Doya K. Modulators of decision making. *Nat Neurosci* 2008; **11:** 410–16.
31 Lidow MS, Goldman-Rakic PS, Gallager DW, Rakic P. Distribution of dopaminergic receptors in the primate cerebral cortex: quantitative autoradiographic analysis using [3H]raclopride, [3H] spiperone and [3H]SCH23390. *Neuroscience* 1991; **40:** 657–71.
32 Humphries MD, Wood R, Gurney K. Dopamine-modulated dynamic cell assemblies generated by the GABAergic striatal microcircuit. *Neural Netw* 2009; **22:** 1174–88.
33 Jiang J, Summerfield C, Egner T. Attention sharpens the distinction between expected and unexpected percepts in the visual brain. *J Neurosci* 2013; **33:** 18438–47.
34 Frank MJ, Scheres A, Sherman SJ. Understanding decision-making deficits in neurological conditions: insights from models of natural action selection. *Philos Trans R Soc Lond B Biol Sci* 2007; **362:** 1641–54.
35 Cisek P. Cortical mechanisms of action selection: the affordance competition hypothesis. *Philos Trans R Soc Lond B Biol Sci* 2007; **362:** 1585–99.
36 Friston KJ, Shiner T, FitzGerald T, et al. Dopamine, affordance and active inference. *PLoS Comput Biol* 2012; **8:** e1002327.
37 Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. The computational anatomy of psychosis. *Front Psychiatry* 2013; **4:** 47.
38 Seth AK, Suzuki K, Critchley HD. An interoceptive predictive coding model of conscious presence. *Front Psychol* 2011; **2:** 395.

39 Seth AK. Interoceptive inference, emotion, and the embodied self. *Trends Cogn Sci* 2013; **17:** 565–73.

40 Ainley V, Tajadura-Jiménez A, Fotopoulou A, Tsakiris M. Looking into myself: changes in interoceptive sensitivity during mirror self-observation. *Psychophysiology* 2012; **49:** 1504–08.

41 Critchley H, Seth A. Will studies of macaque insula reveal the neural mechanisms of self-awareness? *Neuron* 2012; **74:** 423–26.

42 Panksepp J, Solms M. What is neuropsychoanalysis? Clinically relevant studies of the minded brain. *Trends Cogn Sci* 2012; **16:** 6–8.

43 Hoffman RE. Neural network simulations, cortical connectivity, and schizophrenic psychosis. *MD Comput* 1997; **14:** 200–08.

44 Stephan KE, Friston KJ, Frith CD. Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. *Schizophr Bull* 2009; **35:** 509–27.

45 Cohen JD, Servan-Schreiber D. Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychol Rev* 1992; **99:** 45–77.

46 Winterer G, Weinberger DR. Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends Neurosci* 2004; **27:** 683–90.

47 Laruelle M. Schizophrenia: from dopaminergic to glutamatergic interventions. *Curr Opin Pharmacol* 2014; **14:** 97–102.

48 Lisman JE, Coyle JT, Green RW, et al. Circuit-based framework for understanding neurotransmitter and risk gene interactions in schizophrenia. *Trends Neurosci* 2008; **31:** 234–42.

49 Gonzalez-Burgos G, Lewis DA. NMDA receptor hypofunction, parvalbumin-positive neurons, and cortical gamma oscillations in schizophrenia. *Schizophr Bull* 2012; **38:** 950–57.

50 Lisman J. Excitation, inhibition, local oscillations, or large-scale loops: what causes the symptoms of schizophrenia? *Curr Opin Neurobiol* 2012; **22:** 537–44.

51 Jardri R, Denève S. Circular inferences in schizophrenia. *Brain* 2013; **136:** 3227–41.

52 Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. *Science* 1997; **275:** 1593–99.

53 Kishida KT, Montague PR. Imaging models of valuation during social interaction in humans. *Biol Psychiatry* 2012; **72:** 93–100.

54 Dayan P, Daw ND. Decision theory, reinforcement learning, and the brain. *Cogn Affect Behav Neurosci* 2008; **8:** 429–53.

55 O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 2004; **304:** 452–54.

56 Fiorillo CD, Tobler PN, Schultz W. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 2003; **299:** 1898–902.

57 Aron AR, Shohamy D, Clark J, Myers C, Gluck MA, Poldrack RA. Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *J Neurophysiol* 2004; **92:** 1144–52.

58 Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF. Neural systems responding to degrees of uncertainty in human decision-making. *Science* 2005; **310:** 1680–83.

59 Montague PR, Hyman SE, Cohen JD. Computational roles for dopamine in behavioural control. *Nature* 2004; **431:** 760–67.

60 Schultz W, Preuschoff K, Camerer C, et al. Explicit neural signals reflecting reward uncertainty. *Philos Trans R Soc Lond B Biol Sci* 2008; **363:** 3801–11.

61 Moustafa AA, Sherman SJ, Frank MJ. A dopaminergic basis for working memory, learning and attentional shifting in Parkinsonism. *Neuropsychologia* 2008; **46:** 3144–56.

62 Friston K, Schwartenbeck P, Fitzgerald T, Moutoussis M, Behrens T, Dolan RJ. The anatomy of choice: active inference and agency. *Front Hum Neurosci* 2013; **7:** 598.

63 Huys QJ, Dayan P. A Bayesian formulation of behavioral control. *Cognition* 2009; **113:** 314–28.

64 Kapur S. Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am J Psychiatry* 2003; **160:** 13–23.

65 Fleming SM, Dolan RJ, Frith CD. Metacognition: computation, biology and function. *Philos Trans R Soc Lond B Biol Sci* 2012; **367:** 1280–86.

66 Fleming SM, Maloney LT, Daw ND. The irrationality of categorical perception. *J Neurosci* 2013; **33:** 19060–70.

67 Yoshida W, Dziobek I, Kliemann D, Heekeren HR, Friston KJ, Dolan RJ. Cooperation and heterogeneity of the autistic mind. *J Neurosci* 2010; **30:** 8815–18.

68 Kishida KT, King-Casas B, Montague PR. Neuroeconomic approaches to mental disorders. *Neuron* 2010; **67:** 543–54.

69 Moutoussis M, Fearon P, El-Deredy W, Dolan RJ, Friston KJ. Bayesian inferences about the self (and others): a review. *Conscious Cogn* 2014; **25:** 67–76.

70 Deco G, Jirsa VK. Ongoing cortical activity at rest: criticality, multistability, and ghost attractors. *J Neurosci* 2012; **32:** 3366–75.

71 Haimovici A, Tagliazucchi E, Balenzuela P, Chialvo DR. Brain organization into resting state networks emerges at criticality on a model of the human connectome. *Phys Rev Lett* 2013; **110:** 178101.

72 Kitzbichler MG, Smith ML, Christensen SR, Bullmore E. Broadband criticality of human brain network synchronization. *PLOS Comput Biol* 2009; **5:** e1000314.

73 Phillips WA, Silverstein SM. Convergence of biological and psychological perspectives on cognitive coordination in schizophrenia. *Behav Brain Sci* 2003; **26:** 65–82, discussion 82–137.

74 Lynall ME, Bassett DS, Kerwin R, et al. Functional connectivity and brain networks in schizophrenia. *J Neurosci* 2010; **30:** 9477–87.

75 Deco G, Rolls ET, Romo R. Synaptic dynamics and decision making. *Proc Natl Acad Sci USA* 2010; **107:** 7545–49.

76 Friston K, Breakspear M, Deco G. Perception and self-organized instability. *Front Comput Neurosci* 2012; **6:** 44.

77 Rolls ET, Loh M, Deco G, Winterer G. Computational models of schizophrenia and dopamine modulation in the prefrontal cortex. *Nat Rev Neurosci* 2008; **9:** 696–709.

78 Peled A. Neuroscientific psychiatric diagnosis. *Med Hypotheses* 2009; **73:** 220–29.

79 Edwards MJ, Adams RA, Brown H, Pareés I, Friston KJ. A Bayesian account of 'hysteria'. *Brain* 2012; **135:** 3495–512.

80 Fletcher PC, Frith CD. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci* 2009; **10:** 48–58.

81 Pellicano E, Burr D. When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends Cogn Sci* 2012; **16:** 504–10.

82 Mathys C, Daunizeau J, Friston KJ, Stephan KE. A bayesian foundation for individual learning under uncertainty. *Front Hum Neurosci* 2011; **5:** 39.

83 Adams RA, Perrinet LU, Friston K. Smooth pursuit and visual occlusion: active inference and oculomotor control in schizophrenia. *PLoS One* 2012; **7:** e47502.

84 Brodersen KH, Deserno L, Schlagenhauf F, et al. Dissecting psychiatric spectrum disorders by generative embedding. *Neuroimage Clin* 2013; **4:** 98–111.

85 Garrido MI, Kilner JM, Kiebel SJ, Stephan KE, Baldeweg T, Friston KJ. Repetition suppression and plasticity in the human brain. *Neuroimage* 2009; **48:** 269–79.

86 Moran RJ, Symmonds M, Stephan KE, Friston KJ, Dolan RJ. An in vivo assay of synaptic function mediating human cognition. *Curr Biol* 2011; **21:** 1320–25.

87 Moran RJ, Jung F, Kumagai T, et al. Dynamic causal models and physiological inference: a validation study using isoflurane anaesthesia in rodents. *PLoS One* 2011; **6:** e22790.

88 Shergill SS, Samson G, Bays PM, Frith CD, Wolpert DM. Evidence for sensory prediction deficits in schizophrenia. *Am J Psychiatry* 2005; **162:** 2384–86.

89 Shergill SS, White TP, Joyce DW, Bays PM, Wolpert DM, Frith CD. Functional magnetic resonance imaging of impaired sensory prediction in schizophrenia. *JAMA Psychiatry* 2014; **71:** 28–35.

90 Fogelson N, Litvak V, Peled A, Fernandez-del-Olmo M, Friston K. The functional anatomy of schizophrenia: a DCM study of predictive coding. *Schizophr Res* (in press).

91 Bastos-Leite AJ, Ridgway GR, Silveira C, Norton A, Reis S, Friston KJ. Dysconnectivity within the default mode in first-episode schizophrenia: a stochastic dynamic causal modeling study with functional magnetic resonance imaging. *Schizophr Bull* 2014; published online June 17. DOI:10.1093/schbul/sbu080.

92   Skewes JC, Jegindø EM, Gebauer L. Perceptual inference and autistic traits. *Autism* 2014; published online Feb 12. DOI: 10.1177/1362361313519872.

93   Gómez C, Lizier JT, Schaum M, et al. Reduced predictable information in brain signals in autism spectrum disorder. *Front Neuroinform* 2014; **8:** 9.

94   Happé F, Frith U. The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *J Autism Dev Disord* 2006; **36:** 5–25.

95   Paton B, Hohwy J, Enticott PG. The rubber hand illusion reveals proprioceptive and sensorimotor differences in autism spectrum disorders. *J Autism Dev Disord* 2012; **42:** 1870–83.

96   Gu X, Hof PR, Friston KJ, Fan J. Anterior insular cortex and emotional awareness. *J Comp Neurol* 2013; **521:** 3371–88.

97   Eilam-Stock T, Xu P, Cao M, et al. Abnormal autonomic and associated brain activities during rest in autism spectrum disorder. *Brain* 2014; **137:** 153–71.

98   Joyce DW, Averbeck BB, Frith CD, Shergill SS. Examining belief and confidence in schizophrenia. *Psychol Med* 2013; **43:** 2327–38.

99   Averbeck BB, Evans S, Chouhan V, Bristow E, Shergill SS. Probabilistic learning and inference in schizophrenia. *Schizophr Res* 2011; **127:** 115–22.

100 Moutoussis M, Bentall RP, El-Deredy W, Dayan P. Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. *Cogn Neuropsychiatry* 2011; **16:** 422–47.