NEUROPSYCHOPHARMACOLOGY REVIEWS
Advances in the computational understanding of mental illness

Quentin J. M. Huys [1,2], Michael Browning[3,4], Martin P. Paulus[5] and Michael J. Frank [6,7]

Computational psychiatry is a rapidly growing field attempting to translate advances in computational neuroscience and machine learning into improved outcomes for patients suffering from mental illness. It encompasses both data-driven and theory-driven efforts. Here, recent advances in theory-driven work are reviewed. We argue that the brain is a computational organ. As such, an understanding of the illnesses arising from it will require a computational framework. The review divides work up into three theoretical approaches that have deep mathematical connections: dynamical systems, Bayesian inference and reinforcement learning. We discuss both general and specific challenges for the field, and suggest ways forward.

INTRODUCTION
Understanding the main function of an organ is important when attempting to either understand the syndromes it engenders, or when attempting to treat them. For instance, a theoretically grounded understanding of the heart's pumping function allows us to relate shortness of breath to altered pressure gradients. The distinctive function of the brain is computation. The brain processes information, and alters how it processes information as a function of the information it has processed in the past ('learning'). That is, the brain uses information as the currency to make models of the world in order to maximize short- and long-term adaptation to the environment. As such, using computation to turn information into models and to extract information from models is the quintessential function that needs to be understood. The basic premise of computational psychiatry is that alterations in the computations it performs can lead to its malfunction - mental illness. In fact, this view suggests that computational 'errors' can lead to illness in the absence of any other 'neural' problems, and can even lead to illness as a function of past computations or processed information.

Computational psychiatry views illnesses and symptoms through a computational lens. As an example, consider perceptual disturbances. Perception depends strongly on the disambiguation of ambiguous and noisy sensory information through the integration with other information previously acquired. This integration process can be formalized as a probabilistic inference process. Doing so allows the perceptual disturbances to be characterized and linked to specific underlying processes, and thereby also to the underlying biology.

Prior to proceeding, we emphasize that illnesses are complex phenomena defying simplistic etiological or mechanistic accounts [1]. They are likely pluralistic and multi-causal involving multiple levels [2]. Indeed, research has identified contributions to the syndromes we identify as disorders arising at different levels from genetics to neural circuits, psychological processes, and social or societal factors. From a broad computational view, illness arises when a mismatch occurs between the brain's computational ability and the environmental or situational demands placed upon it. For instance, alterations in learning from positive and negative decision outcomes, due to an imbalance in corticostriatal dopaminergic function, can cause either impulsivity such as pathological gambling [3], or tenacity in the face of frequent setbacks. Whether the imbalance results in a feature or a problem depends on background factors such as which goals one has in the first place (i.e., what counts as positive or negative: the reward function), and the statistics of rewards and losses.

Computational investigations are often subdivided into three levels [4]. At the most conceptual level, a computational understanding answers questions about what problem the system solves. For instance, what is the problem in finding actions which are good in the longer term? And precisely why should this be done? At a more concrete level, computational models are of algorithmic nature and describe what computations can be used to achieve a particular goal. Finally, computational models can concern the implementation of algorithms. These three levels are in principle independent. However, the strength of computational modeling is that it allows the connections across these levels to be made – and may even be necessary for doing so. More generally, explanatory accounts of psychiatric disorders need to integrate across biological, psychological and social-environmental domains with inherent many to many relationships [5]. The argument proposed here is that a broad computational approach is useful – and maybe even necessary – to do this in a quantitative manner.

The explanatory models we focus on here are often 'generative' [6–9], meaning that the models can be run on the experiment that individuals were subjected to and generate comparable data. This has the advantage that the explanatory scope of models can be rigorously tested by comparing data generated by the model with observed experimental data. Furthermore, they do so through the latent manipulation of variables that capture computational

[1]Division of Psychiatry and Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK; [2]Camden and Islington NHS Trust, London, UK; [3]Computational Psychiatry Lab, Department of Psychiatry, University of Oxford, Oxford, UK; [4]Oxford Health NHS Trust, Oxford, UK; [5]Laureate Institute For Brain Research (LIBR), Tulsa, OK, USA; [6]Cognitive, Linguistic & Psychological Sciences, Neuroscience Graduate Program, Brown University, Providence, RI, USA and [7]Carney Center for Computational Brain Science, Carney Institute for Brain Science Psychiatry and Human Behavior, Brown University, Providence, RI, USA
Correspondence: Quentin J. M. Huys (q.huys@ucl.ac.uk)

4

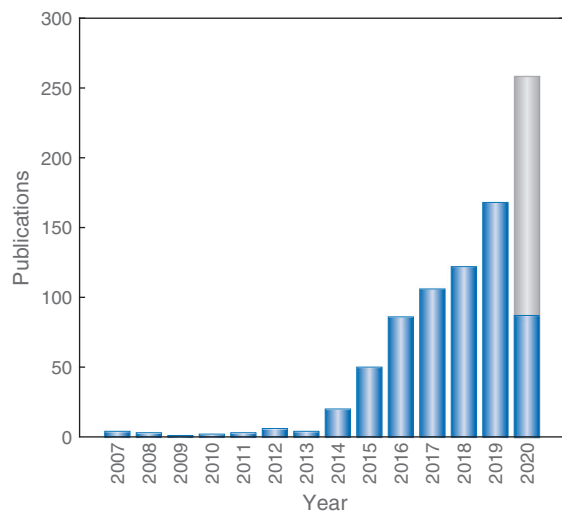**Fig. 1** **Count of publications listed on pubmed and referring to "computational psychiatry" in title, abstract or keywords.** The grey bar is a linear extrapolation for 2020 based on citations up to 1 May 2020.

processes. As such, these modeling approaches are a way to quantitatively test complex but detailed hypotheses about mental processes (c.f. [10–12]). This is in contrast to descriptive models, which may describe the statistical properties of data correctly, but whose internal machinations are less directly interpretable or informative about the underlying mechanisms.

Both theory- and data-driven computational approaches to psychiatric illness are developing rapidly. While previous snapshot summaries of the area exist [8, 13–22], the rate of progress and the number publications in the field (Fig. 1) both mean that the state of the art rapidly moves past these, and that even in a substantive review as here it will not be possible to review all the work.

The contributions in this volume highlight many facets of *data-driven* evidence-based and empirical computational work. These are rapidly advancing the field and allowing researchers and clinicians to deal with – and put to good use – the deluge of data facing them. In the present contribution, we will summarize what we view as the most important recent advances in *theory-driven* computational work relevant to psychiatry [23]. Our aim is two-fold: first, to illustrate, through examples, the usefulness of theory-driven computational approaches for understanding mechanisms in psychiatric illnesses. Second, to provide an updated snapshot of the field. We structure the paper in terms of the class of computational technique used. Briefly, the brain is a dynamical system, and has to solve two fundamental problems: it has to deal with (irreducible) uncertainty, and it has to exert control to survive. As such, we start with dynamical systems, and then turn to Bayesian inference and reinforcement learning. We begin each section with a (very!) brief intuitive summary of that technique and then highlight the important work completed using it over the last few years. We end the review with a summary and synthesis of the progress made, the challenges the field faces and the key next steps.

We note that both inference and learning can be seen as special instances of dynamical systems [24, 25], while in certain situations learning and inference are two faces of the same coin [26, 27]. As such, there are deep mathematical connections between dynamical systems, learning and inference.

DYNAMICAL MODELS
Mental illness can be conceptualized as a dynamic process, i.e., a state of being that changes over time. Dynamical models focus on

Box 1: Dynamical systems mathematical concepts

Differential and difference equation
These equations take the general form of $d\mathbf{x}/dt = f(\mathbf{x})$, where $d\mathbf{x}/dt$ is the rate at which a variable changes. By relating functions f of variables $\mathbf{x}$ to the rate at which the variables change, such equations capture how variables evolve over time. For instance, if perceived palpitations in panic attacks increase proportional to perceived palpitations, then they will grow very rapidly. Difference equations are the analog of differential equations in discrete time.

Attractor dynamics
Some differential equations will evolve towards a set of values (the attractor) when started within a certain range of values (the basin of attraction). In Fig. 3a, for instance, the attractor is the set of neural activations which form a bump in space.

Fixed point
A fixed point is a particular value $\mathbf{x}$ where the system does not change, meaning that it will remain at that point. Fixed points can be *attracting* if the system converges to the fixed point when started within a basin of attraction around the fixed point, akin to a ball rolling to the bottom of a valley; or they can be *unstable* if they diverge away from that point unless they are exactly at that point, akin to a ball on top of a mountain.

Limit cycle
A limit cycle is a particular type of attractor, where the system does not settle on one particular point, but rather cycles through a (possibly arbitrarily long) repetitive trajectory. Neural action potentials are limit cycles.

Excitation-inhibition (E/I) balance
Excitation and inhibition in neural tissues need to be finely balanced at multiple scales to allow for a stable range of dynamic phenomena. Alterations to this balance has profound impacts on the observed dynamics of neural networks.

Divisive normalization
This refers to a particular type of inhibition of neurons, whereby the activity of a neuron is divided by the local pooled activity surrounding the index neuron. Divisive normalization accounts for neural contrast invariance and other receptive field features in primary and higher cortices.

the rules that govern the changes of states over time, the response to environmental input and the emerging consequences from these rules. The dynamics involve an interaction of multiple factors, and these interactions can unfold in surprising and complex ways over time. For instance, the textbook view of panic attacks involves a positive feedback cycle, where interoceptive signals such as palpitations augment anxiety, which in turn increases arousal and ventilatory rate resulting in increased interoceptive signals. A positive feedback cycle is one particular dynamical phenomenon which can be displayed by biological, physical, societal and other dynamical systems. The field of dynamical systems theory is concerned with the mathematical characterization of dynamical systems, which are sets of differential equations (or difference equations if in discrete time) describing how variables interact with and influence each other over time (Fig. 2). Such equations can give rise to numerous dynamical phenomena such as attractors, oscillations, phase transitions and even chaos ([28]; see Box 1).

An important insight arising from the study of dynamical systems is that the observed behavior of a system is often *independent* of the nature of the components involved (Boxes 2 and 3). That is, the same dynamical phenomena can be observed at many different levels and can describe populations of neurons [29], symptoms within an individual [30, 31], interactions between individuals or groups of individuals [32]. Independently of the nature of the unit, the relative behavior of the units will be determined by the dynamical parameters, and they will display phenomena such as attractor states, oscillations and other more complex dynamical phenomena. Furthermore, and maybe most importantly, the behavior of a dynamical system will often determine the overall trajectory of the system, at the same time as being completely at odds with the behavior of the individual components which make up the system.

Hence, the assertion that mental illnesses are dynamic [33–35] has profound and potentially far-reaching consequences: it may be impossible to understand the evolution of psychiatric symptoms without understanding the complex interactions joining them together. A consequence of this complexity is that the
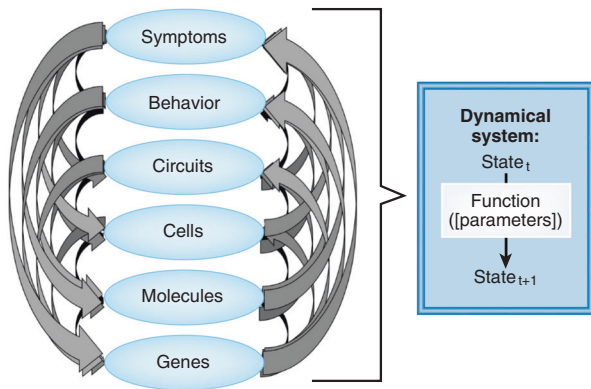
Advances in the computational understanding of mental illness
QJM Huys et al.

5

Fig. 2 Dynamical Systems. Multiple interactions between phenomena at different levelscan result in behavior that is described by a dynamical system, which is guided by a function that takes the state of the system at time t to some later time $t + 1$, with parameters that may transcend any individual level but can be observed across levels of analyses.

effect of interventions which alter the state of some components of the system, such as pharmacotherapy or psychotherapy, will depend on the current global state of that system. In other words, as frequently observed in clinical practice, the same treatment administered to the same patient at different times will produce different, and often counterintuitive effects if the state of the patient has changed. Focusing attention on one's breathing may be good when calm, but exacerbate the panic attack during the attack. Reflecting their general nature, dynamical systems approaches have provided numerous insights into phenomena ranging from intracellular to societal scales.

Linking cellular to cognitive and circuit processes via dynamical systems
Applications of dynamical systems approach at the circuit level have combined algorithmic and biophysical components, and thereby allowed an understanding of how alterations at the cellular or subcellular level impact the ability of circuits to perform certain functions. Particular attention has been paid to so-called attractor dynamics. One type of attractor is a stable point, where the activity of each unit remains approximately constant over time and returns to this activation level if slightly perturbed. Such stable attractors have been extensively studied as network models for persistent neural activities in working memory [36–38], including for their ability to maintain a continuous quantity, e.g., a spatial location [39]. Because the stability of the overall network activity pattern depends on their dynamic properties and how the units interact, such models can be used to examine how properties at the cellular level such as dopamine [40], serotonin [41–43] or NMDA receptor function [44, 45] affect the dynamics, and in turn how they affect the ability of the network to retain information.

Indeed, detailed predictions from such a model were shown to capture working memory sensitivity to distractors in schizophrenia [46], while also accounting for the effects of ketamine [45]. Briefly, in this model NMDA receptors on interneurons affect the extent to which neurons inhibit their neighbors. This in turn affects the profile of the stable attractor 'bump' in the network. A reduction in the efficacy of the NMDA receptor leads to a broadening and reduced stability of the bump, and thereby to an increase in sensitivity to distractors (Fig. 3a). Critically, both the broadening and the increase in distractibility could be demonstrated empirically and shown to be correlated (Fig. 3b; [46]). Indeed, direct recordings of frontal neural assemblies in two animal models of schizophrenia - chronic ketamine administration and a 22q11.2 deletion

model - show direct evidence of impaired attractor stability [47], and a related model has recently been shown to explain disruptions in serial dependencies in working memory in schizophrenia and NMDA receptor encephalitis [48]. The notion of a change in attractor properties in schizophrenia has also been proposed in algorithmic models of decision-making [49], which have also suggested specific relationships of to positive and negative symptoms [50]. Circuit-level attractor dynamics can be put to various uses for computational purposes, most classically for pattern completion and memory recollection [51, 52], but also for Bayesian inference [53, 54], multisensory integration [55] and decision-making more generally [56]. As such, these computational models allow a number of higher-level cognitive functions to be related to mechanistic details at the cellular level, and thus afford an understanding of various psychiatric disturbances [21].

Taking a step back, attractor dynamics are one type of computation dependent on the broader principle of balanced excitation and inhibition (E/I). Alterations to E/I balance have been suggested in a number of other illnesses, in particular also in Autism Spectrum Disorders (ASD) [57]. Computational models of E/I imbalance in ASD have focused on a feature of local circuitry called divisive normalization. This is a very widespread computation important for gain adaptation in visual and auditory primary cortices [58, 59], and alterations in models of divisive normalization can account for alterations in visual (c.f. Figure 3) and auditory perception and possibly also higher cognitive functions in autism [60–64]. Divisive normalization has also been suggested as one way of implementing marginalization in neural circuits [65]. Marginalization is a key step in probabilistic inference algorithms (e.g., belief propagation, c.f. Box 1), and this provides one link for how alterations in local E/I balance could have pervasive impacts on many different cognitive functions, and particularly so on functions requiring appropriately dealing with unknown latent variables such as the intentions of others.

At a larger scale, one notable application of dynamical models to interactions between brain areas capitalized on the fact that the dynamical properties of a system depend on the interaction between its components, and therefore alterations in one component can be counteracted by alterations in another [35, 43]. To understand how serotonergic medication might alleviate glutamatergic deficits, [66] built a spiking network model in which cognitive dorsolateral prefrontal cortical (dlPFC) and affective ventral anterior cingulate (vACC) areas had reciprocal inhibitory interactions. Glutamatergic deficits in depression were modeled as a less efficient glutamate clearance, leading to a situation where the affective vACC was hyperactive and impaired the cognitive dlPFC performance through its excitatory projections to dlPFC inhibitory interneurones. Serotonergic medication in the model was effective at treating this by hyperpolarizing the excitatory vACC cells via 5HT-1A receptors.

The dynamical systems applications to psychiatry reviewed so far are relatively detailed, and have mostly been used to give qualitative accounts of experimentally observed phenomena in the sense that they have not (with very few exceptions [67, 68]) directly been fitted to experimental data. As such their strength lies in their ability to qualitatively link biophysical and cellular details to higher-level phenomena.

Dynamical systems for causality, prediction and control
A different approach has been to fit dynamical systems to time-series data relevant to mental health. There is a very rich literature concerned with efficient identification of dynamical features of systems from multivariate time-series data. Applications include using such dynamical systems models to understand how variables interact; to characterize the overall dynamical characteristics of the system; and to investigate
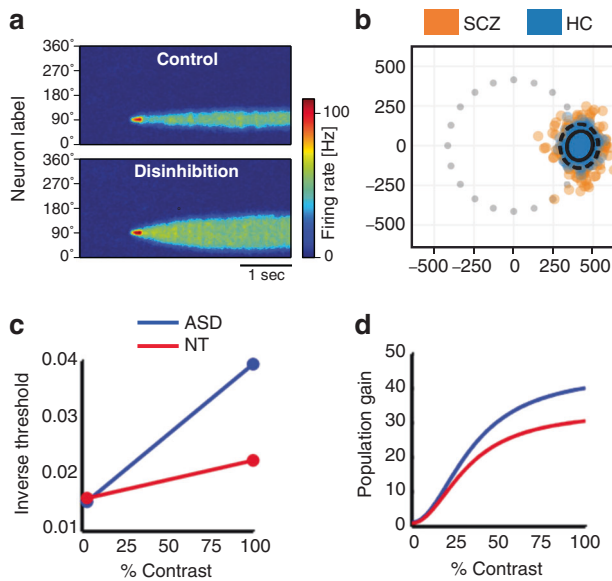
6

Fig. 3 Dynamical system applications. a This shows the activities of a set of neurons, each representing a particular direction after receiving a brief stimulation input at 90 deg. In the control network (top), the population activity is stable and remains tightly focused on the input, hence maintaining this information faithfully. A reduction of the inhibitory input due to a NMDA receptor dysfunction leads to gradually widening bump and hence a less specific representation. **b** In a delayed working memory task, patients with Schizophrenia (orange) show a greater increase in variance in remembered locations over time than healthy controls (blue). **a** and **b** adapted from [46]. **c** Neurotypicals (NT) and patients with autism spectrum disorder (ASD) perform equally well when judging the direction of motion of a low-contrast stimulus, but patients with autism perform better (they have higher inverse thresholds) at higher contrasts. **d** Reducing divisive normalization increases the population gain in the model and allows it to qualitatively capture the improvement seen in ASD. **c** and **d** reproduced from [63].

potential interventions, i.e., to identify how to control the system under study.

Probably the most common approach is in fitting dynamical systems to neural data using autoregressive models to estimate Granger causality, or dynamic causal models (DCM; [69]). These have been used extensively over the past two decades to examine interactions between neural substrates and their breakdown in mental illness. They are distinct from functional connectivity approaches which focus on the correlational structure, in that they imply an underlying model of how the data come about, and allow such models to be explicitly tested [6, 69, 70]. DCM connectivity estimates, for instance, have suggested that the absence of illusions such as the hollow-mask illusion in patients with schizophrenia is due to a reduction in the influence of top-down frontal projections [71, 72].

A very promising direction is the combination of the parameters resulting from such fits with machine-learning techniques for classification of predictive purposes [9, 73, 74]. [75] recently applied this to the relationship between fMRI BOLD responses to emotional faces and the long-term course of depressive disorders. Not only did they find that the connectivity estimates differentiated patients with a good and a poor longitudinal course. But because DCM involves the fitting of an interpretable dynamical system, they were able to then investigate parameters of the fits and point to aberrant (reduced) modulation of the connections within and between the amygdala and face perception areas (fusiform and occipital face areas) by emotions. One major limitation of DCM models has traditionally been the need to fit the model to a few selected areas.

As such, statements about the causality of interactions were subject to confounds due to other areas not included in the analyses. A whole-brain approach has recently been developed [76, 77] which will help to address this. Although it involves important approximations, this promises to bring the same whole-brain causal connectivity approach to fMRI that autoregressive models and Granger causality have previously brought to EEG and MEG [70]. Interesting newer approaches include the examination of controllability in brain networks [78–80], and the use of nonlinear dynamical systems to directly characterize more complex dynamical modes of brain activity [81, 82].

Dynamical systems have also been applied to the analysis of self-report time-series data [83]. Quantitative approaches to within-subject longitudinal data are important. First the temporal course of psychiatric symptoms is an important window to examining the dynamics of mental health disorders. Second, there are fundamental limitations on the ability to generalize from cross-sectional findings to within-subject causes [84, 85]. Third, such an approach implies that symptoms can directly influence each other, or whereby symptoms at least are indicators of processes that can interact directly, and do so independently from any underlying disease process [86, 87]. For instance, sleep disturbances and fatigue are both symptoms of depression in both DSM and ICD [88, 89]. However, it is eminently obvious that sleep disturbances can directly cause daytime fatigue independently of the presence of any depressive disorder. Furthermore, empirical estimates of comorbidity patterns between categorical diagnoses closely track the overlap in symptoms ([86] c.f. [90]).

These observations suggest that patterns of symptoms may inherently stabilize each other, and that hence interactions between symptoms may contribute both to the emergence and stabilization of mental illnesses (e.g., [30, 31]). Indeed, in depression the transition between episodes of wellness and illness shows features of so-called critical slowing-down [91], which is seen when a fixed point becomes unstable and the system transitions into a different stable fixed point. Furthermore, a strong coherence between different symptoms is predictive of a more chronic long-term course of depression [92], possibly because the symptoms maintain each other and stabilize the overall syndrome.

Different types of models have since been applied to self-report time series data, ranging from autoregressive models [93, 94] to linear dynamical systems [95] and highly nonlinear dynamical systems such as Ising models [96, 97] and most recently recurrent neural networks [82]. The relative merits of these approaches are beyond the scope of this review, and there are important questions surrounding the reliability and value of complex measures of such self-report time-series [98]. Nevertheless, a very interesting application of such idiographic research is naturally psychotherapy [99]. For instance, the controllability of a linear dynamical system can be quantified and captures how costly it is to move the system into any target state. Furthermore, such systems can be studied to ask questions such as which symptom is most important in that its alteration would have the strongest desirable impact [100].

INFERENCE: DEALING WITH UNCERTAINTY
Uncertainty is baked into our lives, and two decades of work have shown that the brain pays detailed attention to this [101–103]. Uncertainty plays a transdiagnostic role in most if not all mental illnesses, be it because it is underestimated (e.g., in delusions), overestimated and aversive (e.g., in anxieties) or appetitive (e.g., in certain types of impulsivity).

Mathematically, the most consistent and correct approach to dealing with uncertainty is through Bayes' theorem. This suggests that beliefs should correspond to a distribution over potential explanations h as $p(h)$. This distribution over explanations can be

Advances in the computational understanding of mental illness
QJM Huys et al.

7

Box 2: Inference - mathematical concepts

Probability distribution
A way of describing the probabilities of occurrence of all possible outcomes for some event. In inference, probability distributions can be used to describe beliefs. For example, we could represent the belief of a depressed patient about how enjoyable parties are as a probability for every possible level of enjoyment, from the lowest to the highest. An advantage of using probability distributions as opposed to single numbers to describe beliefs (as used by simple RL algorithms such as the Rescorla–Wagner rule, see section 3) is that distributions capture uncertainty. For example, whereas a single number could describe the belief "Parties are not very enjoyable", a distribution can describe the belief "Parties are probably not very enjoyable, but they might be quite enjoyable".

Joint probability
A joint probability is the probability that two specific outcomes for separate events occurred. The two events might be independent, for example, the probabilities that parties are enjoyable and that the next car I see on the road will be red, or they might be dependent, for example, the probability that parties are enjoyable and that I will enjoy the party happening this evening. The joint probability for outcomes a and b is written as $p(a, b)$.

Conditional probability
A conditional probability is the probability that an outcome will occur, given a separate outcome is known to have happened. For example, the probability that I will enjoy tonight's party given that parties are rarely enjoyable. The conditional probability that event a occurred, given event b is written as $p(a|b)$.

Marginalization
The probability distribution over one variable can be obtained by summing over other variables in a joint distribution. This summation is referred to as marginalization. Two variables are independent if their joint distribution is equal to the product of their marginals. Marginalization is often necessary to properly incorporate the influence of unobserved or latent variables.

Bayes theorem
Bayes theorem describes a simple relationship between conditional probabilities. In inference, it can be used to describe how beliefs should be updated in the face of new experience. For example, how the belief about parties should be revised following the experience of attending another party. The theorem states that the new belief, the posterior, is proportional to the existing belief, called the prior, multiplied by the likelihood (see below for definitions of these terms).

Prior
A belief that some outcome will occur before new evidence has been experienced. For example, this might be the belief that parties are enjoyable before attending a party this evening.

Likelihood
The likelihood is a particular conditional probability, namely the probability of an event given a prior belief. For example, the probability that the party tonight will be enjoyable given that parties are rarely enjoyable.

Posterior
The posterior is a conditional probability describing the same belief as the prior, but following the update of this belief by the evidence. For example, it might describe the belief about how enjoyable parties are, following the experience of attending tonight's party.

---

updated with evidence e by multiplying it with the likelihood function $p(e|h)$ such that the new belief state incorporating the evidence is

$$p(h|e) \propto p(e|h)p(h) \qquad (1)$$

Here, the likelihood term $p(e|h)$ effectively measures how compatible each hypothesis h is with the evidence e, and allows the hypotheses to be weighed by their compatibility with all the evidence or experience. The resulting distribution $p(h|e)$ is the posterior distribution. Further evidence can be included by repeating this step:

$$p(h|e, e_{new}) \propto p(e_{new}|h)p(h|e) \qquad (2)$$

The belief state before including the new evidence $p(h|e)$ was the posterior, but is now the prior. When repeating inference to include new information, the previous conclusion ('posterior') belief becomes the new prior belief. Hence, priors are one of the vehicles through which past experience shapes the interpretation of the present.

Disentangling prior from likelihood biases
A number of mental illnesses are thought to be characterized by particular biases in aspects of prior beliefs affecting inference

about certain experiences or hypotheses. However, estimating priors is often difficult [104]. This is because in controlled experimental situations priors are measured through responses to 'evidence' presented in form of stimuli. This, however, usually means that the experimentally observed responses reflect the posterior $p(h|e)$ rather than the prior, and hence differences between individuals could either be due to differences in the prior $p(h)$ or the likelihood $p(e|h)$ [105].

One way of measuring priors is by examining responses to multiple different stimuli where the stimulus provides ambiguous information and the influence of the prior can hence be estimated as the consistent bias across these stimuli. In depression, prior beliefs are thought to be biased towards hypotheses that make losses or punishments more likely. For instance, when presented with ambiguous information about the probability of obtaining a reward, self-reported optimism covaries with a prior belief in a computational model of choices [106], while depressed patients are less able to update their initial belief when presented with disconfirming positive information [107], possibly due to alterations to anterior cingulate regions [108]. Conversely, anxiety seems to be characterized by a pessimistic bias in the process of accumulating evidence, rather than in the prior [109, 110].

Prior beliefs have been extensively examined in research on psychosis, where alterations in the integration of prior beliefs with evidence have long been postulated to underlie hallucinations and delusions [111, 112]. In this research, the term 'prior' takes on a meaning that is specific to the particular study. While, to date, there is no clear consensus on what the precise pattern of impairments is in psychosis [113], research is progressing rapidly and providing an increasingly nuanced view.

One approach is to examine whether participants automatically infer the statistical regularities in an experiment, and use this to disambiguate information on a given trial. [114] explicitly manipulated this, and found that neither schizotypal nor autistic traits affected it, but that autistic traits were characterized by more accurate perception. The majority of studies have examined more explicitly 'trained' priors. In these, information in a task can be disambiguated through the use of information provided in another form or another part of the task. For instance, [115] binarized images into black and white such that it was very difficult to recognize any shapes in the images. However, when participants had been pre-exposed to the original image in color, they could leverage this prior information and improve their ability to discriminate the images. Strikingly, participants with early-stage psychosis showed a stronger improvement with this prior information. That is, these data suggest an improvement in the ability to integrate these two types of information, and this integration can, but need not, be viewed as the impact of a prior. Similar finding emerge when using visual stimuli as priors on ambiguous auditory stimuli [116], but not when using auditory stimuli to bias ambiguous visual percepts [117].

A related process is the study of perceptual stability with prolonged bistable stimuli. Here, healthy participants who are delusion-prone [118], and patients with schizophrenia [119] show less stable percepts, suggesting that the pure integration of information over time into stable percepts is impaired (c.f. [120]). This weakened low-level information meant that the percept was more easily shaped by providing cues, which could be viewed as stronger sensitivity to trained prior information [118] in the sense of [115]. Finally, [116] used a computational model based on Eq. (2), to examine the trial-by-trial influence from prior trials onto the current trial, and found that it was a stronger in participants with psychosis.

This latter view was recently refined in a new incentivized version of the classic beads task [121] on which participants with psychosis sampled more information, rather than less. A detailed computational (RL; see below) model carefully adjusted for

8

socioeconomic confounds, suggested that patients with delusions formed stronger 'prior' beliefs quickly, and then found it difficult to shift away from these [122]. This appears in principle in keeping with a previous study which suggested that the jumping to conclusion bias was driven by noise in the decision-making process, and also not due to sampling costs [123, 124].

Sequential inference models
Uncertainty is also generated by changes in the world: as the world changes, evidence gathered in the past loses some of its relevance. Some of these changes may be expected, others not. The simplest sequential models simply maintain a running average of experiences, e.g., $h_{t+1} = h_t + a(e_t - h_t)$ (e.g., [125]). Here, the estimation h_t is updated by a fraction of the difference between h_t and the evidence h_t, with the size of this fraction being controlled by the learning rate, α, which is a number between 0 and 1.

An approach that is becoming increasingly popular applies Bayesian inference using latent variable models, such as the Kalman filter or Hidden Markov Models [26, 126] to explicitly estimate sources of uncertainty and effectively adapt the learning rate α over time [127–132]. In these models (Fig. 4a), the true state of the world (the true hypothesis h) changes over time. Such changes can be captured in a very general manner by a distribution $p(h_{t+1}|h_t)$ where the latent state at the next time $t + 1$ depends on the latent state at the current time t. Evidence at time t is now of course directly informative about h_t, but the extent to which it is informative about another time depends on how the world evolves, i.e., on $p(h_{t+1}|h_t)$. As such, the maintenance and discarding of past information implicitly represents an assumption about the stability of the world.

For example, variability caused by changes in the underlying association (sometimes referred to as unexpected uncertainty [133]), means that existing beliefs are less likely to accurately reflect current associations so learners should be more influenced by recent outcomes and use a higher learning rate (see Fig. 4). In contrast, variability caused by random chance (sometimes called expected uncertainty [133]), as occurs in less deterministic associations, reduces how informative each outcome is prompting a reduced learning rate. As such, the learning rate can be viewed as an assumption about how rapidly things change or how random outcomes are.

This development has allowed research which asks whether psychiatric symptoms are associated with the ability to flexibly update beliefs. Evidence suggests impaired updating in anxiety [134, 135], particularly with respect to punishments [136], uncertainty in social interactions [137], and autism [138], with studies in patients with schizophrenia reporting both impaired [116, 139] and excessive updating [49]. A related line of work indicates that humans assume positive and negative outcomes differ in their stability and can adjust these separately [140]. This process allows individuals to treat positive and negative outcomes as if they were differentially informative, providing a potential mechanism for the affective biases believed to be causally related to depression [141–143].

Finally, uncertainty has a kind of value in itself [144]. It is useful to sample uncertain options as this will improve our understanding of them allowing us to make better future choices [145]. However, sampling the unknown can also be hazardous, particularly in aversive environments where novel options are more likely to be dangerous (perhaps leading to negative prior beliefs about the environment as discussed in section 3.1). An active literature has proposed a range of algorithms by which the value of uncertainty may be estimated and used to bias reinforcement learning [144, 146]. In terms of clinical presentations, in anxiety disorders uncertainty appears to be aversive and avoided [147], while in opioid addiction a tolerance of ambiguity is predictive of relapses [148].
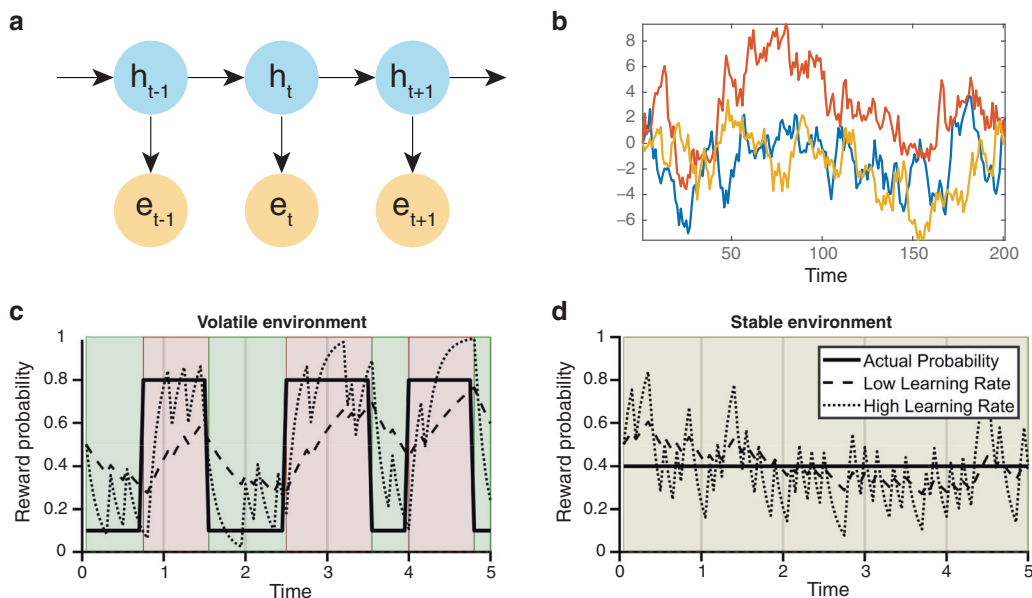

Fig. 4   Learning rates. a Latent variable model with temporal dynamics. Here, the hidden variable h_t evolves according to some dynamics, but is not observed. The observations e_t are directly informative about h at the same timepoint, but the extent to which they are informative about future time points depends on the dynamics of h. **b** Three example trajectories from an AR(1) Ornstein–Uhlenbeck process. This is the evolution of the underlying variable of interest assumed by a Rescorla–Wagner model with a fixed learning rate of 0.05. **c, d** Learning Rates should reflect the changeability of learned associations (Figure adapted from [103]). The reward expectations of learners with a high (dotted line) and low (dashed line) learning rate (LR) are shown in two different environments. Panel **c** illustrates a volatile environment in which the learned association is changing rapidly. As can be seen, the learner with a high learning rate α is better able to update its expectation following changes in association whereas the learner with a low learning rate α never catches up. In contrast, panel **d** illustrates a stable environment in which the learner with the low learning rate accurately estimates the underlying association, with the expectation of the learner with the high learning rate being pulled away from the true value by chance outcomes.

Advances in the computational understanding of mental illness
QJM Huys et al.

9

REINFORCEMENT LEARNING

Symptoms of psychiatric illness very commonly involve alteration of hedonic experience or of behaviors which lead to rewarding or punishing outcomes. This observation has driven interest in how humans learn about rewarding and punishing outcomes and how they use what they have learnt to make decisions. It has also been argued that failure modes in decision-making allow for a principled exploration of dysfunctions on a normative platform [105].

The field of reinforcement learning (RL) is concerned with deriving behavior which maximizes rewards or minimizes losses in the longer term, i.e., not just immediately, but in principle until the end of time. In principle, this is hard, because many things can happen in the future. One of the core insights is that these long-term expectations of future rewards V are governed by a deceptively simple rule:

$$\mathcal{V}(s) = \mathbb{E}[\mathcal{R}(s, s') + \mathcal{V}(s')] \tag{3}$$

This means that the total expected future rewards $\mathcal{V}(s)$ in a state s differ from the total expected future rewards in the next state s' exactly by the amount of reward received on average when going from s to s'. Taking the difference between these two sides provides the temporal reward prediction error signal which can be used to learn the true $\mathcal{V}$ ([149]; see [7] for a very brief introduction).

$$\mathcal{V}_{t+1}(s) = \mathcal{V}_t(s) + \alpha(r_t + \mathcal{V}_t(s') - \mathcal{V}_t(s)) \tag{4}$$

The equation describes how to update the reward expectation of an agent at time t, $\mathcal{V}_t(s)$ in response to experienced outcomes, r_t and a transition to a new state s'. We note that it is similar to the simple running average equation in the previous section, only that here what is averaged is the immediate reward plus the value of the following state. This bootstrapping is at the core of why reinforcement learning can estimate long-term rewards and support optimal decision-making. Strikingly, dopamine neurons appear to report this prediction error with surprising precision [150], and this has fueled an immense research effort, of which we here review only the most recent advances.

In applying Eq. (4) to mental illnesses, a number of questions immediately arise. First the term r_t is supposed to capture both rewards and punishments. Clearly, this is an oversimplification. A further question is how effort should be treated. Second, Eq. (4) effectively describes a type of learning, i.e., of information maintenance. How does this interact with other systems that maintain information, such as working and episodic memory? Third, Eq. (4) does not make reference to knowledge of the world. Clearly, beliefs about how the world work potently influence behavior and learning, and this will be discussed in the section on model-based decision-making. Finally, Eq. (4) makes reference to states s. What are they? A final question is the one discussed in the previous section on what the learning rate α should be.

Reward sensitivity

One of the core symptoms of depression is anhedonia, a reduction in the subjective experience of rewards and motivation. Several studies have shown associations of anhedonia with reduced learning from rewarding outcomes [151–153]. However, dysfunctional reward learning may arise from a number of sources: aberrant updating of values, reduced ability to maintain those values (see section on working memory and RL below), distorted estimates of the reward value of outcomes, or an inability to utilize learned values when selecting actions. Furthermore, learning rate and reward sensitivity trade off each other: in many tasks, it is possible to compensate for any reduction in reward sensitivity by increasing the learning rate. Although there is variability in the literature [152–156], the most consistent effect is that increased

---

Box 3: Reinforcement learning - mathematical concepts

Agent
An agent takes actions, usually with a goal to maximize the sum of its future rewards, given the (sensory or cognitive) state. In computer science, the agent is an algorithm embedded in a computer, but the same notion can be applied to biological agents.

State (S)
A state is the situation the agent is in, which might comprise sensory and cognitive variables. Sometimes the state is "observable", meaning it is perfectly defined by the environment, other times its identity might need to be inferred. For example, whether it is rainy. There are many sensory states that all indicate in different ways that it might be rainy (sight or sound of rain, weather app on a mobile device, or even simply a strong prior that it will rain in London).

Markov Decision Process (MDP)
An MDP is a quintuple $M = \langle S, A, p, r, \gamma \rangle$ with a state space S, an action space A, a transition function mapping how current states transition to future states based on the action $p{:}S \times A \times S$, a reward function mapping which states are considered rewarding $r{:}S \times A \times S \rightarrow R$, and finally a discount γ determining how much more important the immediate future is compared to the distant future.

Policy (π)
The policy is a decision rule that the agent uses to choose an action in the current state.

State value (V)
The expected reward over the long-run, often discounted with future rewards worth less than more immediate ones. $v(t) = \langle \gamma^0 r(t) + \gamma^1 r(t + 1) + \gamma^2 r(t + 2)... \rangle$. $v_\pi(S)$ is the expected value in the current state under the decisions using policy π. Note that while "Reward" is a short-term single signal that may be received in a given state, value is the sum of all discounted rewards you might anticipate from that state in the future.

State-action value (Q)
The expected reward over the long-run of taking an action a in a state s. The state value V is just the Q values averaged over the probability of taking that action, which in turn is given by the policy π.

Structure learning/State Abstraction
In the real world, the relevant state S is often unknown: when deciding whether to get an umbrella or not, many features might be indicative of the rain state, and many others (e.g., music in the background) might be completely irrelevant to this policy. Structure learning is the process of discovering the relevant states of the environment and abstracting them to be as efficient and useful as possible.

Model-free RL
A model-free agent learns a mapping from states to actions based on previously experienced rewards which are "cached", without explicitly representing the likely future states and their available actions.

Model-based RL
A model-based agent explicitly learns the transition function p. When selecting each action, it "plans" by considering future potential outcomes. As such it is more flexible to any changes that may have recently occurred to the reward values of particular states, but is more computationally expensive.

---

anhedonia is associated with a reduced effective value for rewarding outcomes [157–160]. In other words, individuals with higher anhedonia treat rewarding outcomes as if they were less rewarding than those with lower anhedonia (note there is some evidence that bipolarity is associated with the opposite effect; [161]).

The origin of this reduction is not clear. Primary reward sensitivity e.g., to sucrose or smells does not appear reduced, and the reduction appears most clear in more complex 'secondary' rewards such as pleasant visual stimuli [162], suggesting a locus of dysfunction in the construction of derived values [163], but still in the absence of impairments in the *learning* process itself ([164], though see [155]), suggesting a more model-based etiology ([163]; see below). Alternatively, it has been suggested that an individual's mood may interact with their reward sensitivity, biasing estimates of the value of outcomes [165]. Craving, the desire for drugs of abuse in addiction, indeed does have a multiplicative effect on reward values [166], suggesting that a similar process may also be at work in other illnesses. Indeed, there is some evidence of such sequential interactive processes [167], and it has been suggested that it may improve learning in certain situations [165, 168]. However, a miscalibration of the interplay between mood and estimated value may exacerbate the impact of low mood and lead to fluctuations of mood reminiscent of bipolar disorder [165, 169].

Advances in the computational understanding of mental illness
QJM Huys et al.

10

Effort sensitivity

Motivated behavior involves an adaptive integration of costs vs benefits of engaging in physical or cognitive effort. Several decades of research in rodents and now humans has implicated the dopamine system in the energization of motivated behavior for the sake of maximizing rewards [170], an effect that can be captured by computational models of striatal dopamine in balancing this tradeoff [171, 172].

This framework has been further extended to account for decisions to engage in mental as opposed to physical effort – that is, the choice to perform a cognitively difficult task depending on the incentive [173]. Indeed, recent studies showed that baseline striatal dopamine synthesis capacity, as measured by PET, is predictive of individuals' willingness to engage in cognitive effort [174]. Critically, this effect was not simply a shift in overall preference. Rather, a behavioral economic analysis showed that dopamine effects on preference were due to an amplification of the (monetary) benefits, together with a diminution of the subjective costs, of engaging in mental effort, and thus mirrored the impact of striatal dopamine on cost/benefit decisions more broadly [171]. Moreover, stimulant medications that elevate striatal dopamine increased cognitive motivation specifically by altering this cost/benefit ratio, most strongly in those subjects with low baseline levels [174]. This study suggests that the use of stimulant medications in ADHD and in the general population might be better understood not by enhancing the ability, but rather the subjective motivation to engage in cognitive processes, and raise the possibility that such assessments could be useful for predicting treatment outcomes.

Effort cost/benefit calculation is thought to be biased in patients with MDD and schizophrenia patients with negative symptoms, who exert reduced physical effort with increasing incentives [175–177]. But while dopamine might be involved in emphasizing the benefits over costs, other studies suggest that serotonin is specifically related to cost calculations. Indeed, in a randomized trial in healthy participants, those treated over 8 weeks with escitalopram exhibited increased willingness to exert physical effort for monetary incentives, where the impact of 5HT manipulation could be specifically attributed to a reduction in effort cost [178]. Accordingly, it is notable that remitted MDD patients have also been documented to have a larger sensitivity to anticipated effort cost which could be linked to alterations in computational model parameters, and was predictive of relapse after antidepressant treatment discontinuation [175]. Together, these studies suggest that careful assessment of effort-based decision making – by employing paradigms and models designed to disentangle the component valuation processes – may be promising candidates for predicting treatment outcomes resulting from SSRIs or DA manipulations.

If reductions in cognitive effort associated with apathy or lack of motivation are related to cost/benefit computations, can they be ameliorated by simply increasing incentives? Recent findings suggest that this might be feasible. In a model-based learning task, where cognitive strategies are more effortful but can pay off, reductions in cognitive effort linked to a range of subclinical traits were ameliorated by larger monetary incentives, with these incentive effects especially large in participants with depression, anxiety, or sensation seeking [179]. On the other hand, a transdiagnostic factor of compulsivity was related (in a separate study) to a reduction in mental effort avoidance [180].

Finally, we note that exerting effort is only valuable if the actions are (believed to) directly influence outcomes. Depression, for instance, is characterized by helplessness and hopelessness. Computationally, this can be viewed as a belief that actions have a small probability of leading to the desired outcome [181], which in turn strongly affects the value of those actions [182] and hence the extent to which rewards would be able to motivate efforts associated with the actions.

Episodic and working memory interactions with RL

The simple RL update rules in Eq. (4) are a type of memory because the output of the update at one time point maintained and used as the input at the next time point. These processes are increasingly being studied in the context of cognitive and neural processes associated with episodic and working memory.

Episodic/RL interactions. Episodic memories are affectively biased in several disorders, particularly so in depression, where affectively negative memories are easier to recall. A hallmark of episodic memory is its *automatic and incidental* encoding of individual sensory events such that they are bound into an episode [183]. In fact, contemporaneous reward prediction error signals can further enhance memories for related events [184, 185], perhaps via neuromodulation of hippocampus [184].

In healthy controls positive RPEs have a larger benefit for memory encoding than negative RPEs. However, in depression, this bias is reversed [141, 143], providing a computational formalism to explain negative memory biases in depression. We note that the standard Eq. (4) does not include different learning rates for rewards and losses. A recent extension of RL has shown that biases in the learning rate from rewards and losses lead to biased estimates; and that a distribution of learning rates can be used to maintain distributions over values, which can allow for more efficient learning [186].

Episodic disturbances also exist in PTSD, where aversive memories are heavily biased towards traumatic events and are thought not to be integrated with other memories [187]. Such views raise complex questions about what it means to 'integrate' an event memory. One view from RL is that integrating a memory means generalizing the value information from the particular state where it was experienced to others. Thus, in complement to the above description of RL effects on memory; episodic memory contributions can reciprocally influence and augment reinforcement learning and decision making [188]. Indeed, when episodes are sampled from memory, reward-based decisions are biased such that choices are influenced by outcomes linked to that episodic context [189]. Other normative theories suggest that episodic sampling can be useful during offline replay (e.g., during sleep) for prioritizing which events and affective values should be integrated into memory [190]. Such links are promising algorithmic avenues for exploring potential aberrations of memory integration in PTSD.

Finally, RL signals can also regulate decisions about whether a memory should be counted as familiar or not during declarative memory retrieval. Signal detection theory provides a framework to characterize both memory strength (the difference in familiarity between encoded memories and novel ones), and the criterion on the level of memory strength needed to reach a decision as to whether an event is familiar or not. It is now appreciated that this decision criterion is adaptable according to the reinforcement value of previous memory decisions. Indeed, by manipulating reward prediction errors during such a declarative memory task, it was shown that striatal RPEs serve to adapt participants' criterion for judging events as familiar or not, and that this criterion even transferred to other memory tasks such as free recall [191, 192]. These studies and models could provide a basis for studying pathological adaptation of such criterion leading to biases in memory reporting, for example in prodromal schizophrenia states.

Working memory/RL interactions. Working memory can also greatly influence RL processes to augment learning. Much of the RL literature assumes that progressive learning in instrumental tasks relates to striatal dopaminergic function. However, rapid learning in these tasks is strongly supported by prefrontal working memory processes: participants can simply hold in mind recent stimulus-action-outcome associations and use these to improve performance the next time the same stimulus appears. Indeed,
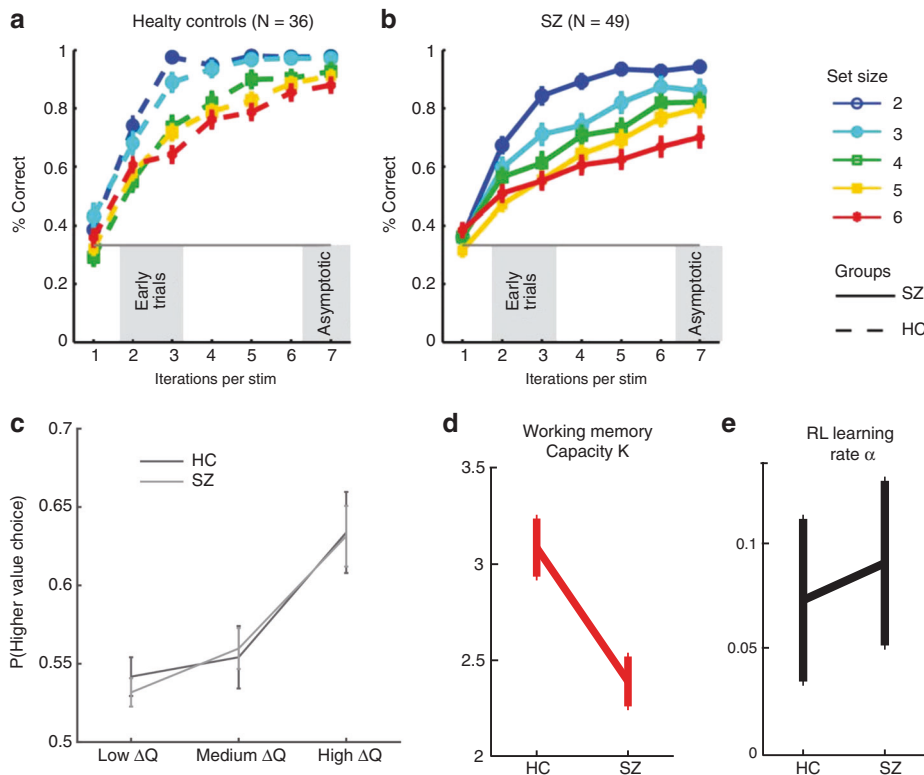
Advances in the computational understanding of mental illness
QJM Huys et al.

11

Fig. 5 Working memory in reinforcement learning. a Working memory effects on reinforcement learning. Learning curves in a simple instrumental learning task are sensitive to the number of stimulus-response contingencies that need to be acquired in the same block (set-size), even given the same number of experiences for each stimulus. These curves are inconsistent with pure RL models but can be captured by an interactive model in which capacity-limited and delay-sensitive working memory processes augment RL, speeding up learning in low set sizes, whereas more incremental but robust RL processes govern asymptotic behavior. **b** Patients with schizophrenia show profound initial learning deficits in these tasks. **c** However, the ability to discriminate amongst stimuli as a function of changes in model-free RL Q values in a transfer test is unimpaired. **d** Modeling isolated the patient deficits during learning to the WM contribution (reduced WM capacity parameter K) whereas RL learning rates α were intact. From [199] and [198].

one of the most celebrated functions of the prefrontal cortex is the active maintenance of information in working memory (WM) in the service of adaptive behavior [193]. Information that is relevant to a current goal is rapidly updated and maintained over time, and can serve to guide upcoming decisions. Interestingly, the degree to which participants engage such top-down processes can affect the net learning rate of an RL system, providing a systems-level mechanism by which distinct brain systems could contribute to the adjustment of learning rate to surprising or uncertain events, as described above.

Of course, these working memory processes are not perfect: they are subject to capacity limitations and sensitive to forgetting, whereas striatal reinforcement learning processes are more incremental but less subject to capacity constraints. This tradeoff provides a normative motivation for the existence of these complementary systems [194] analogous to that described for model-based vs model-free reinforcement learning (see below) [195]. Moreover, it motivates the critical need to consider PFC and WM contributions when evaluating the nature of RL deficits (or the impact of pharmacological treatments) in patient populations. Indeed, many studies report RL deficits in schizophrenia [196, 197], which are sometimes interpreted solely in context of striatal DA-mediated alterations. However, using experimental paradigms and computational models designed to disentangle these processes, reward learning deficits in medicated schizophrenia can be attributed to pronounced reductions in prefrontal WM contributions, with surprisingly intact learning from (and striatal signaling of) reward prediction errors (Fig. 5; [198–200]).

Beyond the independent contributions of these systems, recent studies have further shown that WM and RL processes interact during learning. Top-down WM processes accelerate the acquisition of instrumental contingencies, but because WM can also maintain the expectation that a reward is going to occur, this expectation also reduces the subsequent RPE, as evidenced by both fMRI and EEG [201, 202]. This top-down influence of WM onto RL RPEs is consistent with other observations that cognitive and model-based expectations can modulate model-free RPEs [203, 204] and leads to a counter-intuitive behavioral prediction. Specifically, although instrumental contingencies are learned more rapidly when WM is engaged (i.e., in low load conditions), the reduced RPEs translate to reduced learning in the RL system, leading to more forgetting in the long-term, when WM can no longer be accessed. Conversely, contingencies learned under high WM load are associated with larger RPEs that facilitate plasticity, and are accordingly retained more robustly, as has been shown empirically [198, 202]. This phenomenon may partially explain the surprisingly spared retention of RL contingencies in patients with schizophrenia [198], in spite of profoundly impaired WM during learning, and is concordant with other computational and empirical findings that SZ patients have reduced expected value computations coupled with over-reliance on stimulus-response learning [205]. More speculatively, these findings could imply that other factors that degrade PFC WM processes, such as stress [206], could actually be paradoxically beneficial in terms of long term retention (provided sufficient accumulation of RPEs during initial learning).

Advances in the computational understanding of mental illness
QJM Huys et al.

12

Finally, as in episodic memory, RL signals can also reciprocally influence decisions about WM. That is, how does the brain decide which of several potential pieces of information to store in mind, what to ignore, and when to discard previously maintained information? In models and data, RPEs are used for learning such gating policies in the service of optimizing memory representations that are most useful for task performance [207, 208]. Moreover, in addition to guiding which representations to store, RL strategies are also used to adaptively modulate how much information can be compressed or chunked in working memory [209], an instantiation of the more general notion that many cognitive heuristics or biases can be understood as rational in the face of limited resources [210]. Moreover, this RL-WM interaction provides a coherent set of mechanisms relevant for understanding suboptimal resource allocation, distractibility, and attentional focus, all of which are features of disorders in frontostriatal circuitry including schizophrenia, Parkinson's disease, ADHD, and OCD [211].

Model-based inference: from urges to meaning
Beauty lies in the eye of the beholder, and the meaning or value of events can be profoundly altered in mental illnesses - witness the interpretation of mundane events as profoundly meaningful in delusional mood, or the cognitive distortions characteristic of depression. Formally, how an event influences us in the future depends on what aspects of the event are stored in memory, and how.

The models described so far in Eqs. (4) and (1) are retrospective: Beliefs are purely a function of the past, and the future is expected to behave just as the past did. In both those equations, experiences are evaluated with respect to current reward expectations, used to adapt these, and then discarded. Because α has to be small (to avoid switching after each individual experience), this means that a change in how rewarding an event is will only lead to a change in expected reward after a (large) number of experiences. An alternative approach is to use experiences to build an explicit model of the world, and then use this model to prospectively derive expectations about likely rewards by simulating what might happen in the future [212]. The strength of this model-based approach is that it affords more flexibility to react to a change in how rewarding future events are, but that comes at the computational cost of having to simulate potentially exponentially many future possibilities [195]. While model-based RL is thought to relate to goal-directed decision-making, model-free RL as in Eq. (4) is thought to relate to habitual decisions and incentive salience [195, 204, 213–215].

Several aspects of this model-based/model-free RL distinction are relevant to mental illness. First, and most prominently, a shift away from goal-directed and towards habitual decision-making 'urges' [216] has been demonstrated for OCD [217–220] and associated with prefrontal and myelination impairments [217, 220, 221]. However, this shift is not diagnostically specific, but also extends to a number of other mental illnesses including binge eating, methamphetamine dependence [220] and schizophrenia [222] but not alcohol use disorder [220, 223]. Strikingly, the association is strongest with a 'compulsive' factor extracted across several different questionnaire measures [179, 180, 224–226]. Finally, it appears to have trait features as it does not change with improvement in OCD symptoms [227], even though it is highly sensitive to stress and cognitive load [228–230].

Second, model-based decision-making formalises how beliefs can fundamentally alter how experience influences behavior, i.e., what they 'mean'. This is demonstrated in the now classic experiment by [204]. Depending on the model, a reward can lead to repeating the action that led to it, or it can lead to avoiding it. The avoidance in this case is driven by the interpretation that a different course of action than the one taken can enhance the chances of another reward. Furthermore, full model-based evaluation is so computationally demanding that it only feasible in scenarios that are so simple as to be irrelevant. Hence, goal-directed decision-making is itself subject to a number of approximations, or internal 'decisions' about which aspects of the future to sample [209, 231–233]. Again, these internal decisions must be led by approximate heuristics which can easily result in profound interpretational biases [234]. For instance, discounting temporally distant events relative to proximal ones is a prominent transdiagnostic feature of many illnesses [235, 236], and temporal discounting can be altered by instructing participants to imagine (i.e., to internally simulate) the temporally distant events [237, 238]. As such, a component of temporal discounting may be driven by internal decisions not to simulate or 'think of' certain future events. Similarly, aspects of anxiety [239] and paranoia [124] are thought to relate to model-based assumptions about the future ability to make good choices. Indeed, there are goal-directed components in threat aversion [240], suggesting that imaginary exposure in the psychotherapy of anxiety disorders may act by addressing internal sampling biases [234]. Finally, we note here the relationship between internal simulation decisions and metacognition [226, 241].

Third, however, the experimental ability to distinguish between model-free and model-based decisions depends on the definition of 'states' and 'actions' [242–246], or more generally on the nature of the representation. We will turn to this next.

Structure and abstraction learning
Arguably, for mental illness, we are most interested in the nature of more abstract cognitive *representations* that are used to constrain the state space used for learning and which facilitate transfer and generalization to novel environments. For example, patients with autism exhibit changes in their ability to extract such abstract structure [247].

Model-based processing, in which a person represents the full transition structure of the consequences of their actions on future states, provides one means to be flexible, but is very computationally expensive. While a person could attempt to re-use models from previous contexts, a more efficient strategy is to learn task representations that facilitate re-use of critical components of previous task settings while collapsing over irrelevant aspects, and flexibly recombining bits of learned knowledge to novel situations [248, 249]. Doing so requires aspects of the 'model' to be 'factorized' (that is kept separate from other aspects) while also learning whether such factorization is useful for the given environment. Humans show such flexibility in "generalizing to generalize" which are well captured by such computational considerations [250], but we have yet to understand the mechanisms by which this process occurs, or whether it can be used to understand poor abstraction or aberrant generalization in patients with mental illness.

Generalizing inherently requires learning representations that are compressed: those that retain critical elements of a task or environment structure while discarding details that may not transfer to other situations. While the hippocampus is thought to provide highly pattern-separated conjunctive representations storing specifics, the cortex is thought to provide more elemental and abstract representations [183, 251, 252]. In reinforcement learning, the "successor representation" provides one algorithmic strategy lying in between model-based and model-free learning, which retains aspects of a world model used for planning without all of the specifics [253–258]. Here, a model is represented by considering the impact of the person's actions on the predicted visitation frequency, and reward-predictive values, of future states, without requiring explicit enumeration of each future action and state transition. Mathematically, this is equivalent to learning the predicted sequence of reward prediction errors given the person's actions, while discarding the specific state transitions [254]. Indeed, if one learns abstract structures using only reward-predictive
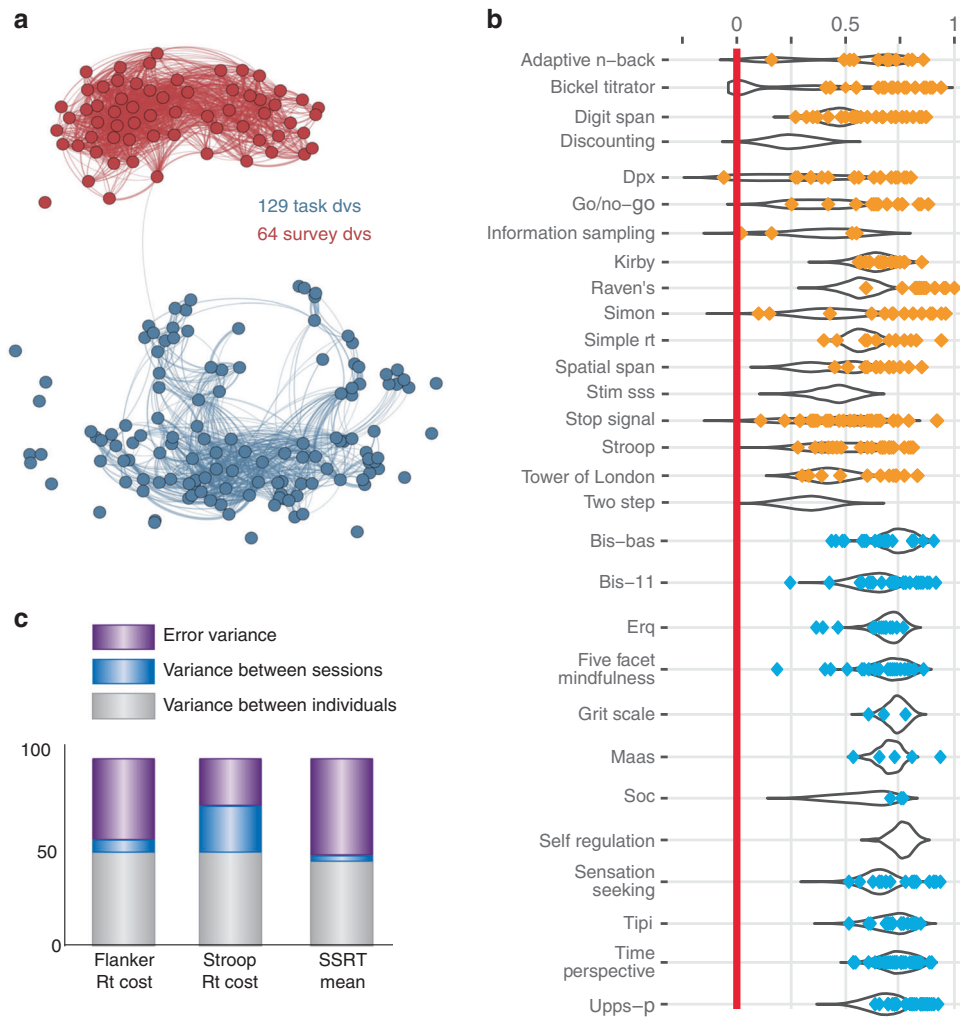
Advances in the computational understanding of mental illness
QJM Huys et al.

13

Fig. 6 Challenges for task-based measurements. a Questionnaire (red) and task-derived measurements (blue) each cohere amongst each other, but there is little coherence between tasks and self-report measurements. Each node represents a measurement, while the edges represent the estimated regularized partial correlation between two measurements. Edges have been thresholded (partial correlation strength ≥ 0.05). From [288]. **b** Meta-analysis and replication of test-retest reliabilities of task- and questionnaire-derived measures in yellow and blue, respectively. Each dot shows a reported test-retest reliability. The black violin plots show results from a large online replication study. Adapted from [281]. **c** Tasks are often designed to show group-level effects, and hence to minimize between-subject variance. However, the low between-subject variance also reduces the scope to see reliable individual differences. The panel shows that only around half the test-retest variance can be accounted for by differences between individuals, meaning that highly robust group-level effects are accompanied by unreliable individual differences. Adapted from [280].

representations, this reduces the dimensionality of the state space in such a way that permits transfer to novel environments that have similar abstract features, even if the specifics of both transitions and rewards change [255]. Critically, such abstract transfer is not afforded by reduced representations that merely maximize reward in the original environment. These computational considerations motivate the study of which brain systems and mechanisms can support learning and re-use of such abstractions and whether they can be fruitfully interrogated to understand the nature of developmental learning disabilities such as ASD.

Pavlovian influences
An important aspect of structure learning relates to a well-established distinction in the animal learning literature, namely the distinction between instrumental conditioning, where reinforcements depend on the animals' behavior, and Pavlovian conditioning, where reinforcements are delivered irrespective of what the animal does. In the latter case, animals (and humans) nevertheless still show behavior - even though it is not necessary. In fact, these

behavioral tendencies are often immutable: animals cannot learn not to salivate when they hear the buzzer. Similar strong Pavlovian tendencies are observable in humans and can profoundly impact decision-making and learning [259, 260]. One way to formally describe these is as a state value which mandates a particular action (e.g., appetitive → approach) [259, 261, 262] and thereby can interfere with instrumental behavior. Alternative possibilities have been considered formally and empirically [263–265].

Pavlovian influences are increased in patients with alcohol use disorder and are predictive of relapse [266] unlike model-based decision-making [223]. Pavlovian escape influences are increased in suicidal patients [267, 268]. In anxiety, there is a subtly different bias towards avoidance behaviors that is independent of Pavlovian values [269].

FUTURE RESEARCH DIRECTIONS
There are at several challenges for computational psychiatry to generate a satisfactory explanatory disease model. First, there is

14

evidence from genetic [270, 271] and circuit level assessments [272, 273] of psychiatric constructs and disorders that there is significant biological and psychological heterogeneity within and across disorders of a similar class. This means that diagnostic labels likely comprise individuals with differing underlying biological architectures. In fact, it has recently been estimated that as much as 80% of polygenic constructs such as anxiety or neuroticism may be due to rare genetic variants that are distributed across the entire genome [274]. The converse, however, is also true: not only does the brain have many ways of producing the same symptoms; the very similar brain dysfunctions can also produce a number of different clinical symptoms. Consider for instance the phenotypic heterogeneity of Huntington's disease. Although as an autosomal dominant disorder it has a simple genetic basis, the clinical variability this results in via the modulation of multiple biochemical pathways is enormous [275]. These clearly are tall orders, and no easy solutions should be expected anytime soon. However, the arguments laid out here suggest that it will be difficult to cut this double Gordian knot without building a computational framework that is able to relate implementational to algorithmic and functional levels.

Second, explanatory variables capture only a small fraction of the observed variance - they do not yet explain enough [276–278]. More specifically, many symptoms of mental illnesses are self-reports expressed in words, and the ability to detect subtle hints in the language of patients is both an important facet of clinicians' skill, but also one that is hard to quantify and hence may contribute to idiosyncrasies and poor agreement between raters. While some of the work reviewed employs causal manipulations that directly alters self-report (e.g., [164]; see also [279]), most of the work reviewed here attempts to gain an understanding of these symptoms through cross-sectional correlations, and these tend to be low even when they replicate robustly [225, 277]. Even if these correlations were high, the guarantees necessary for cross-sectional patterns to be meaningful for individual subjects longitudinally are unlikely to be given [84, 85]. Furthermore, while different, putatively more objective, task-based measures show comparatively better coherence amongst each other, and so do different self-report measures, the coherence between task-derived and self-report measures is relatively poor (Fig. 6a).

Third, part of this is due to an aspect of mechanistic research that was underappreciated until recently, namely the tendency to squash between-subject variability [280]. Although a number of putatively mechanistically informative task-derived measurements are highly robust at the group level, they often show poor test-retest reliability, meaning that individual differences are not robust, and less robust than self-report measures (Fig. 6b; [280–283]). One reason in particular is that group-level effects are maximized when individual differences are minimized. As most mechanistic research employs group-level approaches to discover shared mechanisms, individual variation has often intentionally been suppressed (Fig. 6c). The fitting of generative computational models to data may have an important role to play. Such models can capture multiple aspects of data, such as choices and reaction time, and ensure consistency across all aspects of the data [7]. As such, they can improve the measurement properties by reducing noise and improving test-retest validity (e.g., [284, 285]).

CONCLUSION

Computational psychiatry is a rapidly growing field that combines both data-driven and theory-driven approaches. This review of theory-driven work has shown that investigations into dynamical, inference and learning aspects of mental illnesses are progressing apace and becoming mature. They are allowing increasingly tight relationships between detailed cellular and cognitive processes to be forged and some of these have shown predictive power in longitudinal studies.

As outlined previously [286], a core goal for computational psychiatry is to accelerate the translation of (computational) neuroscience into improved patient outcomes. The paths through which computational methods can support this goal are manifold. First, the focus in this review was on mechanisms. We have illustrated how computational approaches allow mechanistic hypotheses and processes to be tested. In addition, because the brain has a computational function at its heart, they are unavoidable when attempting to grapple with the malfunctions observed in mental illnesses. Second, computational approaches may provide tools for the measurement of these processes, and thereby facilitate precision-psychiatric approaches. For instance, tasks can be used to measure different aspects of learning and inference, and these may be helpful for treatment stratification. Third, the identification of computational processes can motivate novel approaches and interventions. For instance, the work reviewed on the importance of working memory for reinforcement learning in schizophrenia, or on the separate malleability of learning rates for appetitive and aversive events opens up novel potential therapeutic interventions.

Nevertheless, to take this forward, we believe that the field requires a dedicated focus on clinical applications. The field may benefit from a move away from cross-sectional research and towards longitudinal causal or quasi-causal study designs to understand how individuals change over time and respond to interventions.

The cost of acquiring data, and the importance of devising procedures that are robust across labs and indeed across international clinical settings renders relatively large-scale collaborations and consortia critically important [287]. Such collaborations could also be instrumental in setting standards and agreeing on the kinds of details which will make modeling a robust technique for clinical applications.

AUTHOR CONTRIBUTIONS
All authors reviewed literature and jointly wrote the paper.

REFERENCES
1. Kendler KS. Toward a philosophical structure for psychiatry. Am J Psychiatry. 2005;162:433–40.
2. Kendler KS. Explanatory models for psychiatric illness. Am J Psychiatry. 2008;165:695–702.
3. Dagher A, Robbins TW. Personality, addiction, dopamine: Insights from parkinson's disease. Neuron. 2009;61:502–10.
4. Marr D. Vision. New York, NY, USA: Freeman; 1982.

Advances in the computational understanding of mental illness
QJM Huys et al.

15

5. Kendler KS. David skae and his nineteenth century etiologic psychiatric diagnostic system: looking forward by looking back. Mol Psychiatry. 2017;22:802–7.

6. Friston K, Moran R, Seth AK. Analysing connectivity with granger causality and dynamic causal modelling. Curr Opin Neurobiol. 2013;23:172–8.

7. Huys QJM. Bayesian approaches to learning and decision-making. In Anticevic, A and Murray, J, editors, Computational psychiatry: mathematical modelling of mental illness. Elsevier; 2017.

8. Stephan KE, Mathys C. Computational approaches to psychiatry. Curr Opin Neurobiol. 2014;25:85–92.

9. Stephan KE, Schlagenhauf F, Huys QJM, Raman S, Aponte EA, Brodersen KH, et al. Computational neuroimaging strategies for single patient predictions. NeuroImage. 2017;145:180–99.

10. Itani S, Rossignol M, Lecron F, Fortemps P. Towards interpretable machine learning models for diagnosis aid: a case study on attention deficit/hyperactivity disorder. PLoS ONE. 2019;14:e0215720.

11. Liu Y, Admon R, Mellem MS, Belleau EL, Kaiser RH, Clegg R, et al. Machine learning identifies large-scale reward-related activity modulated by dopaminergic enhancement in major depression. Biol Psychiatry: Cogn Neurosci Neuroimaging. 2020;5:163–72.

12. Woo C–W, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci. 2017;20:365–77.

13. Adams RA, Huys QJM, Roiser JP. Computational psychiatry: towards a mathematically informed understanding of mental illness. J Neurol, Neurosurg, Psychiatry. 2016;87:53–63.

14. Corlett PR, Fletcher PC. Computational psychiatry: a rosetta stone linking the brain to mental illness. lancet Psychiatry. 2014;1:399–402.

15. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci. 2016;19:404–13.

16. Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci. 2011;14:154–62.

17. Montague PR. Neuroeconomics: a view from neuroscience. Funct Neurol. 2007;22:219–34.

18. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends Cogn Sci. 2012;16:72–80.

19. Rutledge RB, Chekroud AM, Huys QJ. Machine learning and big data in psychiatry: toward clinical applications. Curr Opin Neurobiol. 2019;55:152–9.

20. Steele JD, Paulus MP. Pragmatic neuroscience for clinical psychiatry. Br J Psychiatry. 2019;215:404–8.

21. Wang X–J, Krystal JH. Computational psychiatry. Neuron. 2014;84:638–54.

22. Wiecki TV, Poland JS, Frank MJ. Model-based cognitive neuroscience approaches to computational psychiatry: clustering and classification. Clin Psychol Sci. 2015;3:378–99.

23. Maia TV, Huys QJM, Frank MJ. Theory-based computational psychiatry. Biol Psychiatry. 2017;82:382–4.

24. Bertsekas DP and Tsitsiklis JN. Neuro-Dynamic Programming. Athena Scientific; 1996.

25. Murphy K, Weiss Y, Jordan MI. Loopy belief propagation for approximate inference: an empirical study. ArXiv, 2013. http://arxiv.org/abs/1301.6725v1.

26. Kalman RE. A new approach to linear filtering and prediction problem. Trans ASME. 1960;82:35–45.

27. Todorov E. General duality between optimal control and estimation. In 2008 47th IEEE Conference on Decision and Control. IEEE; 2008. https://doi.org/10.1109/cdc.2008.4739438.

28. Strogatz SH. Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering. studies in nonlinearity. 2nd edn, Westview Press; 2015.

29. Wang X–J. Synaptic basis of cortical persistent activity: the importance of nmda receptors to working memory. J Neurosci. 1999;19:9587–603.

30. Cramer AOJ, van Borkulo CD, Giltay EJ, van der Maas HLJ, Kendler KS, Scheffer M, et al. Major depression as a complex dynamic system. PloS one. 2016;11: e0167490.

31. Robinaugh DJ, Hoekstra RHA, Toner ER, Borsboom D. The network approach to psychopathology: a review of the literature 2008–18 and an agenda for future research. Psychol Med. 2019;50:353–66.

32. Strawinska-Zanko, U and Liebovitch, LS, (eds) Mathematical modeling of social relationships. Springer International Publishing; 2018. https://doi.org/10.1007/978-3-319-76765-9.

33. Breakspear M. Dynamic models of large-scale brain activity. Nat Neurosci. 2017;20:340–52.

34. Bystritsky A, Nierenberg AA, Feusner JD, Rabinovich M. Computational nonlinear dynamical psychiatry: a new methodological paradigm for diagnosis and course of illness. J Psychiatr Res. 2012;46:428–35.

35. Durstewitz D, Huys QJ, Koppe G. Psychiatric illnesses as disorders of network dynamics. Biological Psychiatry CNNI. 2020. Advance Online Publiation.

36. Amit DJ, Brunel N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb Cortex. 1997;7:237–52.

37. Lisman JE, Fellous J–M, Wang X–J. A role for NMDA-receptor channels in working memory. Nat Neurosci. 1998;1:273–5.

38. Wang M, Yang Y, Wang C–J, Gamo NJ, Jin LE, Mazer JA, et al. NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. Neuron. 2013;77:736–49.

39. Compte A, Brunel N, Goldman-Rakic PS, Wang X–J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb Cortex. 2000;10:910–23.

40. Durstewitz D, Seamans JK, Sejnowski TJ. Neurocomputational models of working memory. Nat Neurosci. 2000;3:1184–91.

41. Cano-Colino M, Almeida R, Compte A. Serotonergic modulation of spatial working memory: predictions from a computational network model. Front Integr Neurosci. 2013;7:71.

42. Cano-Colino M, Almeida R, Gomez-Cabrero D, Artigas F, Compte A. Serotonin regulates performance nonmonotonically in a spatial working memory network. Cereb Cortex (N. Y, N. Y: 1991). 2014;24:2449–63.

43. Maia TV, Cano-Colino M. The role of serotonin in orbitofrontal function and obsessivecompulsive disorder. Clin Psychol Sci. 2015;3:460–82.

44. Cano-Colino M, Compte A. A computational model for spatial working memory deficits in schizophrenia. Pharmacopsychiatry. 2012;45:S49–S56.

45. Murray JD, Anticevic A, Gancsos M, Ichinose M, Corlett PR, Krystal JH, et al. Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. Cereb Cortex. 2014;24:859–72.

46. Starc M, Murray JD, Santamauro N, Savic A, Diehl C, Cho YT, et al. Schizophrenia is associated with a pattern of spatial working memory deficits consistent with cortical disinhibition. Schizophrenia Res. 2017;181:107–16.

47. Hamm JP, Peterka DS, Gogos JA, Yuste R. Altered cortical ensembles in mouse models of schizophrenia. Neuron. 2017;94:153–67.e8.

48. Stein, H, Barbosa, J, Rosa-Justicia, M, Prades, L, Morató, A, Galan, A, et al. (2019). Disrupted serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. 2019, https://www.biorxiv.org/content/10.1101/830471v1.

49. Adams RA, Napier G, Roiser JP, Mathys C, Gilleen J. Attractor-like dynamics in belief updating in schizophrenia. J Neurosci. 2018;38:9471–85.

50. Jardri R, Duverne S, Litvinova A, Deneve S. Experimental evidence for circular inference in schizophrenia. Nat Commun. 2017;8:14218. https://doi.org/10.1038/ncomms14218.

51. Hopfield J. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA. 1982;79:2554.

52. Wills TJ, Lever C, Cacucci F, Burgess N, O'Keefe J. Attractor dynamics in the hippocampal representation of the local environment. Science. 2005;308:873–6.

53. Echeveste R, Aitchison L, Hennequin G, and Lengyel M. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. 2019, https://www.biorxiv.org/content/10.1101/696088v1.

54. Lengyel M, Kwag J, Paulsen O, Dayan P. Matching storage and recall: hippocampal spike timing-dependent plasticity and phase response curves. Nat Neurosci. 2005;8:1677–83.

55. Deneve S, Latham PE, Pouget A. E cient computation and cue integration with noisy population codes. Nat Neurosci 2001;4:826–31.

56. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychol Rev. 2006;113:700–65.

57. Foss-Feig JH, Adkinson BD, Ji JL, Yang G, Srihari VH, McPartland JC, et al. Searching for cross-diagnostic convergence: neural mechanisms governing excitation and inhibition balance in schizophrenia and autism spectrum disorders. Biol Psychiatry. 2017;81:848–61.

58. Carandini M, Heeger DJ. Normalization as a canonical neural computation. Nat Rev Neurosci. 2011;13:51–62.

59. Heeger DJ. Normalization of cell responses in cat striate cortex. Vis Neurosci. 1992;9:181–97.

60. De Martino B, Harrison NA, Knafo S, Bird G, Dolan RJ. Explaining enhanced logical consistency during decision making in autism. J Neurosci: Offcial J Soc Neurosci. 2008;28:10746–50.

61. Lawson RP, Aylward J, White S, Rees G. A striking reduction of simple loudness adaptation in autism. Sci Rep. 2015;5:16157.

62. Louie K, Khaw MW, Glimcher PW. Normalization is a general neural mechanism for context-dependent decision making. Proc Natl Acad Sci USA. 2013;110:6139–44.

63. Rosenberg A, Patterson JS, Angelaki DE. A computational perspective on autism. Proc Natl Acad Sci. 2015;112:9158–65.

Advances in the computational understanding of mental illness
QJM Huys et al.

16

64. Vattikuti S, Chow CC. A computational model for cerebral cortical dysfunction in autism spectrum disorders. Biol Psychiatry. 2010;67:672–8.

65. Beck JM, Latham PE, Pouget A. Marginalization in neural circuits with divisive normalization. J Neurosci: Offcial J Soc Neurosci. 2011;31:15310–9.

66. Ramirez-Mahaluf JP, Compte A. Serotonergic Modulation of Cognition in Prefrontal Cortical Circuits in Major Depression. In: Anticevic, A. & Murray, J. (eds) Computational Psychiatry, Elsevier; 2018. p. 27–46.

67. Moran RJ, Symmonds M, Stephan KE, Friston KJ, Dolan RJ. An in vivo assay of synaptic function mediating human cognition. Curr Biol: CB. 2011;21:1320–5.

68. Symmonds M, Moran CH, Leite MI, Buckley C, Irani SR, Stephan KE, et al. Ion channels in eeg: isolating channel dysfunction in nmda receptor antibody encephalitis. Brain. 2018;141:1691–702.

69. Friston KJ, Harrison L, Penny W. Dynamic causal modelling. Neuroimage. 2003;19:1273–302.

70. Seth AK, Barrett AB, Barnett L. Granger causality analysis in neuroscience and neuroimaging. J Neurosci. 2015;35:3293–7.

71. Dima D, Dietrich DE, Dillo W, Emrich HM. Impaired top-down processes in schizophrenia: a dcm study of erps. NeuroImage. 2010;52:824–32.

72. Dima D, Roiser JP, Dietrich DE, Bonnemann C, Lanfermann H, Emrich HM, et al. Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. NeuroImage. 2009;46:1180–6.

73. Brodersen KH, Deserno L, Schlagenhauf F, Lin Z, Penny WD, Buhmann JM, et al. Dissecting psychiatric spectrum disorders by generative embedding. Neuroimage Clin. 2014;4:98–111.

74. Brodersen KH, Schofield TM, Leff AP, Ong CS, Lomakina EI, Buhmann JM, et al. Generative embedding for model-based classification of fmri data. PLoS Comput Biol. 2011;7:e1002079.

75. Frässle S, Marquand AF, Schmaal L, Dinga R, Veltman DJ, van der Wee NJA, et al. Predicting individual clinical trajectories of depression with generative embedding. NeuroImage Clin. 2020;26:102213.

76. Frässle S, Lomakina EI, Kasper L, Manjaly ZM, Leff A, Pruessmann KP, et al. A generative model of whole-brain effective connectivity. NeuroImage. 2018;179:505–29.

77. Frässle S, Lomakina EI, Razi A, Friston KJ, Buhmann JM, Stephan KE. Regression dcm for fmri. NeuroImage. 2017;155:406–21.

78. Braun U, Schaefer A, Betzel RF, Tost H, Meyer-Lindenberg A, Bassett DS. From maps to multi-dimensional network mechanisms of mental disorders. Neuron. 2018;97:14–31.

79. Gu S, Pasqualetti F, Cieslak M, Telesford QK, Yu AB, Kahn AE, et al. Controllability of structural brain networks. Nat Commun. 2015;6. https://doi.org/10.1038/ncomms9414.

80. Perry A, Roberts G, Mitchell PB, Breakspear M. Connectomics of bipolar disorder: a critical review, and evidence for dynamic instabilities within interoceptive networks. Mol Psychiatry. 2018;24:1296–318.

81. Durstewitz D. A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements. PLoS Comput Biol. 2017;13:e1005542.

82. Koppe G, Toutounji H, Kirsch P, Lis S, Durstewitz D. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fmri. PLoS Comput Biol. 2019;15:e1007263.

83. Piccirillo ML, Rodebaugh TL. Foundations of idiographic methods in psychology and applications for psychotherapy. Clin Psychol Rev. 2019;71:90–100.

84. Borsboom D, Kievit RA, Cervone D, and Hood SB. The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In: Valsiner J, Molenaar PCM, Lyra MCDP, and Chaudhary N, editors. Dynamic process methodology in the social and developmental sciences. Springer Science + Business Media. 2009. https://doi.org/10.1007/978-0-387-95922-1_4.

85. Molenaar PC, Campbell CG. The new person-specific paradigm in psychology. Curr Directions Psychol Sci. 2009;18:112–7.

86. Borsboom D, Cramer AOJ, Schmittmann VD, Epskamp S, Waldorp LJ. The small world of psychopathology. PLoS One. 2011;6:e27407.

87. Fried EI, van Borkulo CD, Cramer AOJ, Boschloo L, Schoevers RA, Borsboom D. Mental disorders as networks of problems: a review of recent insights. Soc Psychiatry Psychiatr Epidemiol. 2016;52:1–10.

88. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM5 R). American Psychiatric Pub; 2013.

89. World Health Organization. International classification of diseases. World Health Organization Press; 1990.

90. Newson JJ, Hunter D, Thiagarajan TC. The heterogeneity of mental health assessment. Front Psychiatry. 2020;11:76.

91. van de Leemput IA, Wichers M, Cramer AOJ, Borsboom D, Tuerlinckx F, Kuppens P, et al. Critical slowing down as early warning for the onset and termination of depression. Proc Natl Acad Sci USA. 2014;111:87–92.

92. van Borkulo C, Boschloo L, Borsboom D, Penninx BWJH, Waldorp LJ, Schoevers RA. Association of symptom network structure with the course of longitudinal depression. JAMA Psychiatry. 2015;72:1219–26.

93. Bringmann LF, Ferrer E, Hamaker EL, Borsboom D, Tuerlinckx F. Modeling nonstationary emotion dynamics in dyads using a time-varying vector-auto-regressive model. Multivar Behav Res. 2018;53:293–314.

94. Bringmann LF, Vissers N, Wichers M, Geschwind N, Kuppens P, Peeters F, et al. A network approach to psychopathology: new insights into clinical longitudinal data. PLoS One. 2013;8:e60188.

95. Lodewyckx T, Tuerlinckx F, Kuppens P, Allen NB, Sheeber L. A hierarchical state space approach to affective dynamics. J Math Psychol. 2011;55:68–83.

96. Loossens T, Mestdagh M, Dejonckheere E, Kuppens P, Tuerlinckx F, Verdonck S. The affective ising model: a computational account of human affect dynamics. PsyArXiv, 2019. https://doi.org/10.31234/osf.io/ky23d.

97. van Borkulo CD, Borsboom D, Epskamp S, Blanken TF, Boschloo L, Schoevers RA, et al. A new method for constructing networks from binary data. Sci Rep. 2014;4:5918.

98. Dejonckheere E, Mestdagh M, Houben M, Rutten I, Sels L, Kuppens P, et al. Complex affect dynamics add limited information to the prediction of psychological well-being. Nat Human Behav. 2019;3:478–91.

99. Molenaar PC. Dynamic assessment and adaptive optimization of the psychotherapeutic process. Behav Assess. 1987;9:389–416.

100. Henry TR, Robinaugh D, Fried EI. On the control of psychological networks. PsyArXiv, 2020. https://doi.org/10.31234/osf.io/7vpz2.

101. Bach DR, Dolan RJ. Knowing how much you don't know: a neural organization of uncertainty estimates. Nat Rev Neurosci. 2012;13:572–86.

102. Doya K, Ishii S, Pouget A, Rao R, editors. Bayesian brain: Probabilistic approaches to neural coding. Cambridge, MA: MIT Press; 2007.

103. Pulcu E, Browning M. The misestimation of uncertainty in affective disorders. Trends Cogn Sci. 2019;23:865–75.

104. Houlsby NMT, Huszár F, Ghassemi MM, Orbán G, Wolpert DM, Lengyel M. Cognitive tomography reveals complex, task-independent mental representations. Curr Biol. 2013;23:2169–15.

105. Huys QJM, Guitart-Masip M, Dolan RJ, Dayan P. Decision-theoretic psychiatry. Clin Psychol Sci. 2015b;3:400–21.

106. Stankevicius A, Huys QJM, Kalra A, Seriès P. Optimism as a prior belief about the probability of future reward. PLoS Comput Biol. 2014;10:e1003605.

107. Rupprechter S, Stankevicius A, Huys QJM, Steele JD, Seriès P. Major depression impairs the use of reward values for decision-making. Sci Rep. 2018;8:13798.

108. Rupprechter S, Stankevicius A, Huys QJM, Series P, Steele JD. Abnormal reward valuation and event-related connectivity in unmedicated major depressive disorder. Psychol Med. 2020. Advance online publication.

109. Aylward J, Hales C, Robinson E, Robinson OJ. Translating a rodent measure of negative bias into humans: the impact of induced anxiety and unmedicated mood and anxiety disorders. Psychological Med. 2020;50:237–46.

110. Kim M, Kim S, Lee K-U, and Jeong B. Pessimistically biased perception in panic disorder during risk learning. Depression Anxiety. 2020.

111. Gray J, Feldon J, Rawlins J, Hemsley D, Smith A. The neuropsychology of schizophrenia. Behav Brain Sci. 1991;14:1–20.

112. Hemsley DR, Garety PA. The formation of maintenance of delusions: a bayesian analysis. Br J Psychiatry. 1986;149:51–6.

113. Sterzer P, Adams RA, Fletcher P, Frith C, Lawrie SM, Muckli L, et al. The predictive coding account of psychosis. Biol Psychiatry. 2018;84:634–43.

114. Karvelis P, Seitz AR, Lawrie SM, Seriès P. Autistic traits, but not schizotypy, predict increased weighting of sensory information in bayesian visual integration. eLife. 2018;7. https://doi.org/10.7554/elife.34115.

115. Teufel C, Subramaniam N, Dobler V, Perez J, Finnemann J, Mehta PR, et al. Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. Proc Natl Acad Sci. 2015;112:13401–6.

116. Powers AR, Mathys C, Corlett PR. Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. Science. 2017;357:596–600.

117. Stuke H, Weilnhammer VA, Sterzer P, Schmack K. Delusion proneness is linked to a reduced usage of prior beliefs in perceptual decisions. Schizophrenia Bull. 2019;45:80–6.

118. Schmack K, G'omez-Carrillo de Castro A, Rothkirch M, Sekutowicz M, Rössler H, Haynes J-D, et al. Delusions and the role of beliefs in perceptual inference. J Neurosci: Offcial J Soc Neurosci. 2013;33:13701–12.

119. Schmack K, Schnack A, Priller J, Sterzer P. Perceptual instability in schizophrenia: probing predictive coding accounts of delusions with ambiguous stimuli. Schizophrenia Res Cognition. 2015;2:72–77.

120. Nour MM, Dahoun T, Schwartenbeck P, Adams RA, FitzGerald THB, Coello C, et al. Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. Proc Natl Acad Sci. 2018;115:E10167–E10176.

Advances in the computational understanding of mental illness
QJM Huys et al.

17

121. Ross RM, McKay R, Coltheart M, Langdon R. Jumping to conclusions about the beads task? a meta-analysis of delusional ideation and data-gathering. Schizophrenia Bull. 2015;41:1183–91.

122. Baker SC, Konova AB, Daw ND, Horga G. A distinct inferential mechanism for delusions in schizophrenia. Brain. 2019;142:1797–812.

123. Ermakova AO, Gileadi N, Knolle F, Justicia A, Anderson D, Fletcher PC, et al. Cost evaluation during decision-making in patients at early stages of psychosis. Comput Psychiatry. 2019;3:18–39.

124. Moutoussis M, Bentall RP, El-Deredy W, Dayan P. Bayesian modelling of jumping-toconclusions bias in delusional patients. Cogn Neuropsychiatry. 2011;16:422–47.

125. Rescorla R and Wagner A. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A and Prokasy, W, editors, Classiacal Conditioning II:Current research and theory. New York : Appleton-Centaury-Crofts; 1972. p. 64–99.

126. Roweis S, Ghahramani Z. A unifying review of linear gaussian models. Neural Comput. 1999;11:305–45.

127. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nat Neurosci. 2007;10:1214–21.

128. Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, et al. Uncertainty in perception and the Hierarchical Gaussian Filter. Frontiers in Human. Front Human Neurosci. 2014;8:825.

129. McGuire JT, Nassar MR, Gold JI, Kable JW. Functionally dissociable influences on learning rate in a dynamic environment. Neuron. 2014;84:870–81.

130. Nassar MR, Bruckner R, Frank MJ. Statistical context dictates the relationship between feedback-related EEG signals and learning. eLife. 2019;8. https://doi.org/10.7554/eLife.46975.

131. Nassar MR, Bruckner R, Gold JI, Li S-C, Heekeren HR, Eppinger B. Age differences in learning emerge from an insu cient representation of uncertainty in older adults. Nat Commun. 2016;7:11609.

132. Nassar MR, Wilson RC, Heasly B, Gold JI. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J Neurosci: Offcial J Soc Neurosci. 2010;30:12366–78.

133. Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. Neuron. 2005;46:681–92.

134. Browning M, Behrens TE, Jocham G, O'Reilly JX, Bishop SJ. Anxious individuals have di culty learning the causal statistics of aversive environments. Nat Neurosci. 2015;18:590–6.

135. Huang H, Thompson W, Paulus MP. Computational dysfunctions in anxiety: failure to differentiate signal from noise. Biol Psychiatry. 2017;82:440–6.

136. Aylward J, Valton V, Ahn W-Y, Bond RL, Dayan P, Roiser JP et al. Altered learning under uncertainty in unmedicated mood and anxiety disorders. Nat. Hum. Behav. 2019.

137. Lamba A, Frank MJ, and FeldmanHall O (2020). Anxiety impedes adaptive social learning under uncertainty. Psychol Sci. 2020; 32343637.

138. Lawson RP, Mathys C, Rees G. Adults with autism overestimate the volatility of the sensory environment. Nat Neurosci. 2017;20:1293–9.

139. Hernaus D, Xu Z, Brown EC, Ruiz R, Frank MJ, Gold JM, et al. Motivational deficits in schizophrenia relate to abnormalities in cortical learning rate signals. Cogn Affective Behav Neurosci. 2018b;18:1338–51.

140. Pulcu E and Browning M. Affective bias as a rational response to the statistics of rewards and punishments. eLife. 2017.

141. Korn CW, Sharot T, Walter H, Heekeren HR, Dolan RJ. Depression is related to an absence of optimistically biased belief updating about future life events. Psychol Med. 2014;44:579–92.

142. Mathews A, MacLeod C. Cognitive vulnerability to emotional disorders. Annu Rev Clin Psychol. 2005;1:167–95.

143. Rouhani N, Niv Y. Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. Psychopharmacology. 2019;236:2425–35.

144. Gershman SJ, Niv Y. Novelty and Inductive Generalization in Human Reinforcement Learning. Top Cogn Sci. 2015;7:391–415.

145. Gittins J, Kevin G, and Richard W. Multi-armed Bandit Allocation Indices. 2nd ed. Hoboken, New Jersey: Wiley; 2011. Library Catalog: www.wiley.com.

146. Schulz E, Gershman SJ. The algorithmic architecture of exploration in the human brain. Curr Opin Neurobiol. 2019;55:7–14.

147. Charpentier CJ, Aylward J, Roiser JP, Robinson OJ. Enhanced risk aversion, but not loss aversion, in unmedicated pathological anxiety. Biol Psychiatry. 2017;81:1014–22.

148. Konova AB, Lopez-Guzman S, Urmanche A, Ross S, Louie K, Rotrosen J et al. Computational markers of risky decision-making for identification of temporal windows of vulnerability to opioid use in a real-world clinical setting. JAMA Psychiatry. 2019;77:368–77.

149. Sutton RS, Barto AG. Reinforcement learning: an introduction. 2nd edn. Cambridge, MA: MIT Press; 2017.

150. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275:1593–9.

151. Eshel N, Roiser JP. Reward and punishment processing in depression. Biol Psychiatry. 2010;68:118–24.

152. Pizzagalli DA, Iosifescu D, Hallett LA, Ratner KG, Fava M. Reduced hedonic capacity in major depressive disorder: evidence from a probabilistic reward task. J Psychiatr Res. 2008;43:76–87.

153. Pizzagalli DA, Jahn AL, O'Shea JP. Toward an objective characterization of an anhedonic phenotype: a signal-detection approach. Biol Psychiatry. 2005;57:319–27.

154. Chase HW, Frank MJ, Michael A, Bullmore ET, Sahakian BJ, Robbins TW. Approach and avoidance learning in patients with major depression and healthy controls: relation to anhedonia. Psychol Med. 2009;40:433–40.

155. Kumar P, Waiter G, Ahearn T, Milders M, Reid I, Steele JD. Abnormal temporal difference reward-learning signals in major depression. Brain. 2008;131(Pt 8):2084–93.

156. Must A, Szabo Z, Bodi N, Szasz A, Janka Z, Keri S. Sensitivity to reward and punishment and the prefrontal cortex in major depression. J Affective Disord. 2006;90:209–15.

157. Cavanagh JF, Bismark AW, Frank MJ, Allen JJB. Multiple dissociations between comorbid depression and anxiety on reward and punishment processing: Evidence from computationally informed EEG. Computational Psychiatry. 2019;3:1–17.

158. Huys QJM, Pizzagalli DA, Bogdan R, Dayan P. Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. Biol Mood Anxiety Disord. 2013;3:12.

159. Lawlor VM, Webb CA, Wiecki TV, Frank MJ, Trivedi M, Pizzagalli DA et al. Dissecting the impact of depression on decision-making. Psychol Med. 2019;1–10.

160. Webb CA, Dillon DG, Pechtel P, Goer FK, Murray L, Huys QJM, et al. Neural correlates of three promising endophenotypes of depression: evidence from the embarc study. Neuropsychopharmacology. 2016;41:454–63.

161. Linke JO, Koppe G, Scholz V, Kanske P, Durstewitz D, Wessa M. Aberrant probabilistic reinforcement learning in first-degree relatives of individuals with bipolar disorder. J Affective Disord. 2020;264:400–6.

162. Bylsma LM, Morris BH, Rottenberg J. A meta-analysis of emotional reactivity in major depressive disorder. Clin Psychol Rev. 2008;28:676–91.

163. Huys QJM, Dayan P, Daw. Depression: a decision-theoretic account. Ann Rev Neurosci 2015a;38:1–23.

164. Rutledge RB, Moutoussis M, Smittenaar P, Zeidman P, Taylor T, Hrynkiewicz L, et al. Association of neural and emotional impacts of reward prediction errors with major depression. JAMA Psychiatry. 2017;74:790–7.

165. Eldar E, Rutledge RB, Dolan RJ, Niv Y. Mood as representation of momentum. Trends Cogn Sci. 2016;20:15–24.

166. Konova AB, Louie K, Glimcher PW. The computational form of craving is a selective multiplication of economic value. Proc Natl Acad Sci. 2018;115:4122–7.

167. Neville V, Dayan P, Gilchrist ID, Paul ES, Mendl M. Dissecting the links between reward and loss, decision-making, and self-reported affect using a computational approach. PsyArXiv. 2020. https://doi.org/10.31234/osf.io/ndc7h.

168. Eldar E, Niv Y. Interaction between emotional state and learning underlies mood instability. Nat Commun. 2015;6:6149.

169. Mason L, Eldar E, Rutledge RB. Mood instability and reward dysregulation—a neurocomputational model of bipolar disorder. JAMA Psychiatry. 2017;74:1275.

170. Salamone JD, Pardo M, Yohn SE, López-Cruz L, SanMiguel N, Correa M. Mesolimbic dopamine and the regulation of motivated behavior. Curr Top Behav Neurosci. 2016;27:231–57.

171. Collins AGE, Frank MJ. Opponent actor learning (opal): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol Rev. 2014;121:337–66.

172. Niv Y, Daw ND, Joel D, Dayan P. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacol (Berl). 2007;191:507–20.

173. Westbrook A, Braver TS. Dopamine does double duty in motivating cognitive effort. Neuron. 2016;91:708.

174. Westbrook JA, van den Bosch R, Maatta JI, Hofmans L, Papadopetraki D, Cools R, et al. Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *co-senior authors. Science. 2020;367:1362–6.

175. Berwian IM, Wenzel JG, Collins AGE, Seifritz E, Stephan KE, Walter H et al. Computational mechanisms of effort and reward decisions in patients with depression and their association with relapse after antidepressant discontinuation. JAMA Psychiatry. 2020.

176. Gold JM, Waltz JW, Frank MJ. Effort cost computation in schizophrenia: a commentary on the recent literature. Biol Psychiatry. 2015;78:747–53.

177. Treadway MT, Bossaller NA, Shelton RC, Zald DH. Effort-based decision-making in major depressive disorder: a translational model of motivational anhedonia. J Abnorm Psychol. 2012;121:553–8.

Advances in the computational understanding of mental illness
QJM Huys et al.

18

178. Meyniel F, Goodwin GM, Deakin JW, Klinge C, MacFadyen C, Milligan H, et al. A specific role for serotonin in overcoming effort cost. eLife. 2016;5.

179. Patzelt EH, Kool W, Millner AJ, Gershman SJ. Incentives boost model-based control across a range of severity on several psychiatric constructs. Biol Psychiatry. 2019a;85:425–33.

180. Patzelt EH, Kool W, Millner AJ, Gershman SJ. The transdiagnostic structure of mental effort avoidance. Sci Rep. 2019b;9:1689.

181. Maier S, Seligman M. Learned helplessness: theory and evidence. J Exp Psychol: Gen. 1976;105:3–46.

182. Huys QJM, Dayan P. A Bayesian formulation of behavioral control. Cognition. 2009;113:314–28.

183. O'Reilly RC, Rudy JW. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. Psychol Rev. 2001;108:311–45.

184. Davidow JY, Foerde K, Galván A, Shohamy D. An upside to reward sensitivity: the hippocampus supports enhanced reinforcement learning in adolescence. Neuron. 2016;92:93–9.

185. Jang AI, Nassar MN, Dillon DG, Frank MJ. Positive reward prediction errors during decision making strengthen memory encoding. Nat Hum Behav. 2019;3:719–32.

186. Dabney W, Kurth-Nelson Z, Uchida N, Starkweather CK, Hassabis D, Munos R, et al. A distributional code for value in dopamine-based reinforcement learning. Nature. 2020;577:671–5.

187. Ehlers A, Clark DM. A cognitive model of posttraumatic stress disorder. Behav Res Ther. 2000;38:319–45.

188. Gershman SJ, Daw ND. Reinforcement learning and episodic memory in humans and animals: an integrative framework. Annu Rev Psychol. 2017;68:101–28.

189. Bornstein AM, Norman KA. Reinstated episodic context guides sampling-based decisions for reward. Nat Neurosci. 2017;20:997–1003.

190. Mattar MG, Daw ND. Prioritized memory access explains planning and hippocampal replay. Nat Neurosci. 2018;21:1609–17.

191. Scimeca JM, Badre D. Striatal contributions to declarative memory retrieval. Neuron. 2012;75:380–92.

192. Scimeca JM, Katzman PL, Badre D. Striatal prediction errors support dynamic control of declarative memory decisions. Nat Commun. 2016;7:1–15.

193. Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annu Rev Neurosci. 2001;24:167–202.

194. Collins AGE, Frank MJ. How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. Eur J Neurosci. 2012;35:1024–35.

195. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005;8:1704–11.

196. Schlagenhauf F, Huys QJM, Deserno L, Rapp MA, Beck A, Heinze H-J, et al. Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. Neuroimage. 2014;89:171–80.

197. Waltz JA, Gold JM. Probabilistic reversal learning impairments in schizophrenia: further evidence of orbitofrontal dysfunction. Schizophr Res. 2007;93:296–303.

198. Collins AGE, Albrecht MA, Waltz JA, Gold JM, Frank MJ. Interactions among working memory, reinforcement learning, and effort in value-based choice: A new paradigm and selective deficits in schizophrenia. Biol psychiatry. 2017a;82:431–9.

199. Collins AGE, Brown, J, Gold J, Waltz J, Frank MJ. Working memory contributions to reinforcement learning in schizophrenia. J Neurosci. 2014;34:13747–56.

200. Dowd EC, Frank MJ, Collins AGE, Gold JM, Barch DM. Probabilistic reinforcement learning in patients with schizophrenia: Relationships to anhedonia and avolition. Biol Psychiatry: Cogn Neurosci Neuroimaging. 2016;1:460–73.

201. Collins AGE, Ciullo B, Frank MJ, Badre D. Working memory load strengthens reward prediction errors. J Neurosci. 2017b;37:4332–42.

202. Collins AGE, Frank MJ. Within- and across-trial dynamics of human eeg reveal cooperative interplay between reinforcement learning and working memory. Proc Natl Acad Sci. 2018;115:2502–7.

203. Collins AGE, Frank MJ. Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. Cognition. 2016;152:160–9.

204. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011;69:1204–15.

205. Hernaus D, Gold JM, Waltz JA, Frank MJ. Impaired expected value computations coupled with overreliance on stimulus-response learning in schizophrenia. Biol Psychiatry: Cogn Neurosci neuroimaging. 2018a;3:916–26.

206. Schwabe L, Wolf OT. Stress prompts habit behavior in humans. J Neurosci. 2009;29:7191–8.

207. Lloyd K, Becker N, Jones M, Bogacz R. Learning to use working memory: a reinforcement learning gating model of rule acquisition in rats. Front Comput Neurosci. 2012;6:87.

208. O'Reilly RC, Frank MJ. Making working memory work: a computational model of learning in the frontal cortex and basal ganglia. Neural Comput. 2006;18:283–328.

209. Nassar MR, Helmers J, Frank MJ. Chunking as a rational strategy for lossy data compression in visual working memory. Psychol Rev. 2018;125:486–511.

210. Lieder F, Gri ths TL. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. Behav Brain Sci. 2020;43:e1.

211. Cools R. Chemistry of the adaptive mind: lessons from dopamine. Neuron. 2019;104:113–31.

212. Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND. Model-based choices involve prospective neural activity. Nat Neurosci. 2015;18:767–72.

213. Huys QJM, Tobler PN, Hasler G, Flagel SB. The role of learning-related dopamine signals in addiction vulnerability. Prog Brain Res. 2014;211:31–77.

214. McClure SM, Daw ND, Montague PR. A computational substrate for incentive salience. TINS. 2003;26:423–8.

215. Schad DJ, Rapp MA, Garbusow M, Nebe S, Sebold M, Obst E, et al. Dissociating neural learning signals in human sign- and goal-trackers. Nat Hum Behav. 2020;4:201–14.

216. Robbins TW, Gillan CM, Smith DG, de Wit S, Ersche KD. Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. Trends Cogn Sci. 2012;16:81–91.

217. Gillan CM, Apergis-Schoute AM, Morein-Zamir S, Urcelay GP, Sule A, Fineberg NA, et al. Functional neuroimaging of avoidance habits in obsessivecompulsive disorder. Am J Psychiatry. 2015;172:284–93.

218. Gillan CM, Morein-Zamir S, Urcelay GP, Sule A, Voon V, Apergis-Schoute AM, et al. Enhanced avoidance habits in obsessive-compulsive disorder. Biol Psychiatry. 2014;75:631–8.

219. Gillan CM, Papmeyer M, Morein-Zamir S, Sahakian BJ, Fineberg NA, Robbins TW, et al. Disruption in the balance between goal-directed behavior and habit learning in obsessivecompulsive disorder. Am J Psychiatry. 2011;168:718–26.

220. Voon V, Derbyshire K, Rück C, Irvine MA, Worbe Y, Enander J, et al. Disorders of compulsivity: a common bias towards learning habits. Mol Psychiatry. 2015;20:345–52.

221. Ziegler G, Hauser TU, Moutoussis M, Bullmore ET, Goodyer IM, Fonagy P, et al. Compulsivity and impulsivity traits linked to attenuated developmental frontostriatal myelination trajectories. Nat Neurosci. 2019;22:992–9.

222. Culbreth AJ, Westbrook A, Daw ND, Botvinick M, Barch DM. Reduced model-based decision-making in schizophrenia. J Abnorm Psychol. 2016;125:777–87.

223. Nebe S, Kroemer NB, Schad DJ, Bernhardt N, Sebold M, Müller DK, et al. No association of goal-directed and habitual control with alcohol consumption in young adults. Addiction Biol. 2018;23:379–93.

224. Gillan CM, Kalanthroff E, Evans M, Weingarden HM, Jacoby RJ, Gershkovich M, et al. Comparison of the association between goal-directed planning and self-reported compulsivity vs obsessive-compulsive disorder diagnosis. JAMA Psychiatry. 2019;1–10.

225. Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. Elife. 2016;5.

226. Rouault M, Seow T, Gillan CM, Fleming SM. Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. Biol Psychiatry. 2018;84:443–51.

227. Wheaton MG, Gillan CM, Simpson HB. Does cognitive-behavioral therapy affect goal-directed planning in obsessive-compulsive disorder? Psychiatry Res. 2019;273:94–99.

228. Otto AR, Gershman SJ, Markman AB, Daw ND. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. Psychol Sci. 2013a;24:751–61.

229. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci USA. 2013b;110:20941–6.

230. Schad DJ, Jünger E, Sebold M, Garbusow M, Bernhardt N, Javadi AH, et al. Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. Front Psychol. 2014;5:1450.

231. Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP. Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol. 2012;8:e1002410.

232. Huys QJM, Lally N, Faulkner P, Eshel N, Seifritz E, Gershman SJ, et al. Interplay of approximate planning strategies. Proc Natl Acad Sci USA. 2015c;112:3098–103.

233. Lally N, Huys QJM, Eshel N, Faulkner P, Dayan P, Roiser JP. The neural basis of aversive pavlovian guidance during planning. J Neurosci: Offcial J Soc Neurosci. 2017;37:10215–29.

234. Huys QJM, Renz D. A formal valuation framework for emotions and their control. Biol Psychiatry. 2017;82:413–20.

Advances in the computational understanding of mental illness
QJM Huys et al.

19

235. Amlung M, Marsden E, Holshausen K, Morris V, Patel H, Vedelago L, et al. Delay discounting as a transdiagnostic process in psychiatric disorders. JAMA Psychiatry. 2019;76:1176.

236. Story GW, Moutoussis M, and Dolan RJ. A computational analysis of aberrant delay discounting in psychiatric disorders. Front Psychol. 2016;6.

237. Hakimi S, Hare TA. Enhanced neural responses to imagined primary rewards predict reduced monetary temporal discounting. J Neurosci. 2015;35:13103–9.

238. Kurth-Nelson Z, Bickel W, Redish AD. A theoretical account of cognitive effects in delay discounting. Eur J Neurosci. 2012;35:1052–64.

239. Zorowitz S, Momennejad I, Daw ND. Anxiety, avoidance, and sequential evaluation. Comput Psychiatry. 2020;4:1–17.

240. Korn CW, Bach DR. Minimizing threat via heuristic and optimal policies recruits hippocampus and medial prefrontal cortex. Nat Hum Behav. 2019;3:733–45.

241. Hauser TU, Allen M, Purg N, Moutoussis M, Rees G, and Dolan RJ. Noradrenaline blockade specifically enhances metacognitive performance. eLife. 2017;6.

242. Daw ND and Dayan P. The algorithmic anatomy of model-based evaluation. Philos Trans R Soc Lond B Biol Sci. 2014; 369(1655).

243. Deserno L and Hauser TU. Beyond a cognitive dichotomy: can multiple decision systems prove useful to distinguish compulsive and impulsive symptom dimensions? Biological Psychiatry. 2020.

244. Dezfouli A, Balleine BW. Habits, action sequences and reinforcement learning. Eur J Neurosci. 2012;35:1036–51.

245. Dezfouli A, Balleine BW. Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. PLoS Comput Biol. 2013;9:e1003364.

246. Shahar N, Moran R, Hauser TU, Kievit RA, McNamee D, Moutoussis M, et al. Credit assignment to state-independent task representations and its relationship with model-based decision making. Proc Natl Acad Sci. 2019b;116:15871–6.

247. Rajendran G, Mitchell P. Cognitive theories of autism. Dev Rev. 2007;27:224–60.

248. Franklin NT, Frank MJ. Compositional clustering in task structure learning. PLOS Comput Biol. 2018;14:e1006116.

249. Wingate D, Diuk C, Donnell T, Tenenbaum J, Gershman S. Compositional policy priors. MIT CSAIL Technical Report 2013-007. 2013.

250. Franklin NT, Frank MJ. Generalizing to generalize: humans flexibly switch between compositional and conjunctive structures during reinforcement learning. PLoS Comput. Biol. 2020;16:e1007720.

251. Behrens TE, Muller TH, Whittington JC, Mark S, Baram AB, Stachenfeld KL, et al. What is a cognitive map? organizing knowledge for flexible behavior. Neuron. 2018;100:490–509.

252. Whittington JC, Muller TH, Mark S, Chen G, Barry C, Burgess N, et al. The Tolman-Eichenbaum machine: unifying space and relational memory through generalisation in the hippocampal formation. 2019. https://www.biorxiv.org/content/10.1101/770495v1.

253. Barreto A, Dabney W, Munos R, Hunt JJ, Schaul T, van Hasselt HP. et al. Successor Features for Transfer in Reinforcement Learning. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, (eds). Advances in Neural Information Processing Systems 30, Curran Associates, Inc.; 2017. p. 4055–65.

254. Lehnert L and Littman ML. Successor features combine elements of model-free and model-based reinforcement learning. arXiv. 2019. doi: 1901.11437.

255. Lehnert L, Littman ML, Frank MJ. Reward-predictive representations generalize across tasks in reinforcement learning. bioRxiv. 2020. https://doi.org/10.1101/653493v2.

256. Momennejad I, Russek EM, Cheong JH, Botvinick MM, Daw ND, Gershman SJ. The successor representation in human reinforcement learning. Nat Hum Behav. 2017;1:680–92.

257. Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND. Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Comput Biol. 2017;13:e1005768.

258. Stachenfeld KL, Botvinick MM, Gershman SJ. The hippocampus as a predictive map. Nat Neurosci. 2017;20:1643.

259. Guitart-Masip M, Huys QJM, Fuentemilla L, Dayan P, Duzel E, Dolan RJ. Go and no-go learning in reward and punishment: interactions between affect and effect. Neuroimage. 2012;62:154–66.

260. Huys QJM, Cools R, Gölzer M, Friedel E, Heinz A, Dolan RJ, Dayan P. Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. PLoS Comput Biol. 2011;7:e1002028.

261. Boureau Y-L, Dayan P. Opponency revisited: competition and cooperation between dopamine and serotonin. Neuropsychopharmacology. 2011;36:74–97.

262. Dayan P, Niv Y, Seymour B, Daw ND. The misbehavior of value and the discipline of the will. Neural Netw. 2006;19:1153–60.

263. Cartoni E, Puglisi-Allegra S, Baldassarre G. The three principles of action: a pavlovianinstrumental transfer hypothesis. Front Behav Neurosci. 2013;7:153.

264. Dorfman HM and Gershman SJ. Controllability governs the balance between pavlovian and instrumental action selection. Nat Commun. 2019;10.

265. Swart JC, Froböse MI, Cook JL, Geurts DE, Frank MJ, Cools R et al. Catecholaminergic challenge uncovers distinct pavlovian and instrumental mechanisms of motivated (in)action. eLife. 2017;6.

266. Garbusow M, Nebe S, Sommer C, Kuitunen-Paul S, Sebold M, Schad DJ, et al. Pavlovian-to-instrumental transfer and alcohol consumption in young male social drinkers: behavioral, neural and polygenic correlates. J Clin Med. 2019;8. https://doi.org/10.3390/jcm8081188.

267. Millner AJ, den Ouden HEM, Gershman SJ, Glenn CR, Kearns JC, Bornstein AM, et al. Suicidal thoughts and behaviors are associated with an increased decision-making bias for active responses to escape aversive states. J Abnorm Psychol. 2019;128:106–18.

268. Millner AJ, Gershman SJ, Nock MK, den Ouden HEM. Pavlovian control of escape and avoidance. J Cogn Neurosci. 2018;30:1379–90.

269. Mkrtchian A, Aylward J, Dayan P, Roiser JP, Robinson OJ. Modeling avoidance in mood and anxiety disorders using reinforcement learning. Biol Psychiatry. 2017;82:532–9.

270. Hall LS, Adams MJ, Arnau-Soler A, Clarke TK, Howard DM, Zeng Y, et al. Genome-wide meta-analyses of stratified depression in generation scotland and uk biobank. Transl Psychiatry. 2018;8:9.

271. Smith DJ, Escott-Price V, Davies G, Bailey ME, Colodro-Conde L, Ward J, et al. Genome-wide analysis of over 106 000 individuals identifies 9 neuroticism-associated loci. Mol Psychiatry. 2016;21:749–57.

272. Wolfers T, Beckmann CF, Hoogman M, Buitelaar JK, Franke B, and Marquand AF. Individual differences v. the average patient: mapping the heterogeneity in adhd using normative models. Psychol Med. 2019; 1–10.

273. Wolfers T, Doan NT, Kaufmann T, Alnaes D, Moberget T, Agartz I, et al. Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. JAMA Psychiatry. 2018;75:1146–55.

274. Wainschtein P, Jain DP, Yengo L, Zheng Z, Cupples LA, Shadyab AH, et al. Recovery of trait heritability from whole genome sequence data. 2019, https://www.biorxiv.org/content/10.1101/588020v1.

275. Ross CA, Aylward EH, Wild EJ, Langbehn DR, Long JD, Warner JH, et al. Huntington disease: natural history, biomarkers and prospects for therapeutics. Nat Rev Neurol. 2014;10:204–16.

276. Barch DM, Carter CS, Committee CE. Measurement issues in the use of cognitive neuroscience tasks in drug development for impaired cognition in schizophrenia: a report of the second consensus building conference of the cntrics initiative. Schiz Bull 2008;34:613–8.

277. Gignac GE, Szodorai ET. Effect size guidelines for individual differences researchers. Personal Individ Differences. 2016;102:74–8.

278. Savitz JB, Rauch SL, Drevets WC. Clinical application of brain imaging for the diagnosis of mood disorders: the current state of play. Mol Psychiatry. 2013;18:528–39.

279. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo C-W, Kross E. An fmri-based neurologic signature of physical pain. N. Engl J Med. 2013;368:1388–97.

280. Hedge C, Powell G, Sumner P. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav Res Methods. 2018;50:1166–86.

281. Enkavi AZ, Eisenberg IW, Bissett PG, Mazza GL, MacKinnon DP, Marsch LA, et al. Large-scale analysis of test-retest reliabilities of self-regulation measures. Proc Nat Acad Sci USA. 2019;116:5472–7.

282. Huys QJM. Computational cognitive methods for precision psychiatry. In Williams, L., editor, Neuroscience-informed precision psychiatry. APA; 2020.

283. Rouder JN, Haaf JM. A psychometrics of individual differences in experimental tasks. Psychonomic Bull Rev. 2019;26:452–67.

284. Brown VM, Chen J, Gillan CM, and Price RB. Improving the reliability of computational analyses: model-based planning and its relationship with compulsivity. Biological Psychiatry CNNI. 2020.

285. Shahar N, Hauser TU, Moutoussis M, Moran R, Keramati M, consortium N, et al. Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. PLoS Comput Biol 2019a;15:e1006803.

286. Paulus MP, Huys QJ, Maia TV. A roadmap for the development of applied computational psychiatry. Biol Psychiatry Cogn Neurosci Neuroimaging. 2016;1:386–92.

287. Browning M, Carter CS, Chatham C, Den Ouden H, Gillan CM, Baker JT, et al. Realizing the Clinical Potential of Computational Psychiatry: Report From the Banbury Center Meeting, February 2019. Biol Psychiatry. 2020; 88:e5-e10.

288. Eisenberg IW, Bissett PG, Enkavi AZ, Li J, MacKinnon DP, Marsch LA, et al. Uncovering the structure of self-regulation through data-driven ontology discovery. Nat Commun. 2019;10.