# Computational Psychiatry

**Robb B Rutledge and Rick A Adams**

**Introduction.** Computational psychiatry is a field that applies methods from computational neuroscience to understanding and treating psychiatric disorders. The lack of good animal models for psychiatric disorders hinders the development of new treatments, and there remains a wide gap between the clinical level operated on by psychiatrists as they evaluate and treat patients and the neurobiological level that is the subject of basic research. The goal of computational psychiatry is to bridge this gap, developing links between different levels of description and hence a deeper understanding of the mechanisms underlying psychiatric disorders. The Global Burden of Disease survey reveals that mental illness is an enormous contributor to disability worldwide with 10% of Disability-Adjusted Life Years (years of life lost combined with years lived with disability) attributable to mental illness. While the number of global Disability-Adjusted Life Years for many disorders declined from 1990 to 2010, the number for mental illness increased by 38% over that period, and this burden is expected to increase in the future (Murray et al., 2012). In this chapter, we review some of the main research areas in computational psychiatry, providing examples of the kind of models that are relevant to understanding the behavior of healthy individuals and individuals with psychiatric disorders including depression, schizophrenia, and autism. We do not provide an exhaustive view, and the interested reader is referred to a number of detailed recent review articles that discuss major issues in this emerging field (Adams, Huys, & Roiser, 2015; Corlett & Fletcher, 2014; Dayan, Dolan, Friston, & Montague, 2015; Huys, Moutoussis, & Williams., 2011; Maia & Frank, 2011; Montague, Dolan, Friston, & Dayan, 2012; Stephan & Mathys, 2014; Wang & Krystal, 2014; Wiecki, Poland, & Frank, 2015).

Psychiatric disorders are disorders of the brain, but diagnosis relies on symptoms that do not relate directly to the underlying mechanisms and provide little information as to the best course of treatment for an individual. Two individuals could have the same symptom, like a depressed mood, for entirely different reasons. Furthermore, two individuals with fundamentally different symptoms could be diagnosed with the same disorder. Diagnoses are descriptive categorizations based, for example, on the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5). However, interrater reliability for the DSM-5 is surprisingly poor, particularly for mood and anxiety disorders (Freedman et al., 2013). The National Institute of Mental Health proposed with the Research Domain Criteria (RDoC) project to develop a transdimensional classification system that cuts across current diagnostic categories. The RDoC project would instead be based on specific behaviors and their underlying neural circuits (Cuthbert & Insel, 2013). In this way, future diagnostic systems would reflect current research from fields including psychology, neuroscience, and genetics, and would provide a better way of understanding symptoms and making treatment decisions. Computational methods can support this effort by making the assumptions of theories explicit so that specific predictions can made at different levels of descriptions, from the molecular to the behavioral level. Aberrant learning and decision making are major features of many psychiatric disorders and much research in computational psychiatry has focused on developing mathematical descriptions of learning and decision making for a variety of disorders (Huys, Guitart-Masip, Dolan, & Dayan, 2015; Montague et al., 2012). The neurobiology underlying learning about rewards is now well understood, and can be described by reinforcement learning theory. In this framework, value expectations are updated on the basis of past experience and used to make future decisions. Learning from rewards is one hallmark of adaptive behavior that may be malfunctioning in psychiatric disorders, and reinforcement learning theory has been applied to model behavioral and neural data in many different psychiatric disorders.

## Computational modeling of mood disorders

Mood disorders including major depressive disorder and bipolar disorder are enormously disruptive and carry large costs for society (Simon, 2003). The two major symptoms of depression in the DSM-5 are: 1) a depressed mood as indicated by subjective report or the observation of others, and 2) a decreased interest or pleasure in most activities (referred to as anhedonia). Dysfunction in the neural mechanisms that compute value is proposed to be the source of the aberrant decisions of depressed individuals (Huys et al., 2015). Researchers distinguish between model-free and model-based valuation mechanisms. Model-free mechanisms learn a direct map between utilities (the subjective value of outcomes) and the states and actions that precede them. While model-free learning is computationally efficient and converges in stable environments, it is notably inflexible, which can lead to errors, particularly in dynamic environments. Model-free learning is thought to be driven by reward prediction errors (RPEs) represented by the neurotransmitter dopamine. RPEs are used to update value estimates (e.g., if you get more than you expected, expect more next time) and these value estimates can be used to improve decision making (Rangel, Camerer, & Montague, 2008). Model-based mechanisms learn a model of the environment that captures the probabilistic relationships between states, actions, and utilities. These models allow different possible future courses of action to be simulated. Such simulations are possible in simple environments but are computationally demanding and actual decisions often reflect the influence of both model-free and model-based learning. Reinforcement learning models can be used to generate trial-by-trial predictions about brain activity that relates to model parameters. Error signals reflecting both model-free and model-based learning have been measured in the striatum with functional MRI (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). Some researchers have suggested that depression might be characterized by a reduced ability to learn about rewards and, consistent with that idea, neuroimaging studies have found reduced RPE signals in the striatum in depressed subjects (Gradin et al., 2011), but often little difference in learning between depressed and control subjects. A meta-analysis found that depression reduced reward sensitivity, which was related to anhedonia, but did not affect the rate of learning (Huys, Pizzagalli, Bogdan, & Dayan, 2013). These results hint that the primary cause of depression is not related to a deficit in dopamine-based model-free learning, which might explain why dopaminergic drugs are not typically effective antidepressants.

Selective serotonin reuptake inhibitors (SSRIs) are the most common pharmacological treatment for depression, despite little evidence for the commonly held belief that low serotonin levels are linked to the low mood of depressed individuals (Cowen & Browning, 2015). Serotonin is linked to the processing of aversive stimuli and to punishment-related behavior (Crockett & Cools, 2015). One theory is that serotonin acts an aversive counterpart to dopamine, but recent evidence from optogenetically identified serotonin neurons in the dorsal raphe finds no evidence for this theory, with tonic activity of serotonin neurons representing information about both average reward and average punishment (Cohen, Amoroso, & Uchida, 2015). SSRIs lead to rapid positive shifts in how the brain responses to emotional stimuli (Harmer et al., 2009). Because SSRIs can take many weeks to have antidepressant effects, one possibility is that SSRIs influence emotional responses to events which over time leads to an individual learning new and more positive associations. It should be possible to support this theory by relating SSRI-induced changes in learning to changes in mood dynamics. However, although depression and other mood disorders are evaluated using subjective measures, the subjective feelings related to mood are poorly understood. If emotions are important for adaptive behavior, the subjective feelings associated with emotions should reflect the activity of the same neural circuits that underlie reinforcement learning and valuation mechanisms. Modeling of mood dynamics may provide useful insights into the dysfunction present in some psychiatric disorders.

## The function of mood and its relation to behavior

Subjective conscious experience is colored by affective states like mood. Unlike emotions, moods can be long lasting and need not have a single cause. As such, the study of mood dynamics may be particularly amenable to the use of computational models, which can attempt to dissociate the many influences on mood. Experience-sampling methods have been used to probe affective states as participants go about their daily lives (Csikszentmihalyi & Larson, 1987; Killingsworth & Gilbert, 2010). These methods have recently been adapted to examine momentary changes in happiness (a proxy for mood) during a value-based decision-making task (Brown et al., 2014; Rutledge, Skandali, Dayan, & Dolan, 2014). Subjects made choices between safe and risky options, gaining and losing small amounts of money, and were asked periodically 'How happy are you at this moment?' Happiness fluctuated throughout the task (Figure 1) and depended on RPEs, the difference between experienced and predicted outcomes. The dynamics of happiness were captured by the following model:

$$\mathrm{Happiness(t)} = \mathrm{w}_0 + \mathrm{w}_1 \sum_{\mathrm{j}=1}^{\mathrm{t}} \gamma^{\mathrm{t-j}}\, \mathrm{CR}_\mathrm{j} + \mathrm{w}_2 \sum_{\mathrm{j}=1}^{\mathrm{t}} \gamma^{\mathrm{t-j}}\, \mathrm{EV}_\mathrm{j} + \mathrm{w}_3 \sum_{\mathrm{j}=1}^{\mathrm{t}} \gamma^{\mathrm{t-j}}\, \mathrm{RPE}_\mathrm{j}$$

where t is trial number, $w_0$ is a baseline parameter, and other weights w capture the influence of different event types. The forgetting factor $0 \leq \gamma \leq 1$ is such that more recent events have a larger impact than earlier trials. Experienced and predicted rewards were reflected in CR, EV, and RPE variables: $CR_j$ is the certain reward if chosen on trial j, $EV_j$ is the expected value or average return for the risky option if chosen on trial j, and $RPE_j$ is the RPE on trial j if the risky choice was chosen. Weights for CR, EV, and RPE variables were significantly positive at the group level in multiple lab-based and smartphone-based (n=18,420) experiments (Rutledge et al., 2014). Forgetting factors were such that events that occurred more than 10 trials ago had essentially no impact on happiness. In summary, recent rewards and expectations both impacted happiness. Happiness depended not on how well subjects were doing in the task, but instead whether they were doing better than expected. Blood-oxygen-level dependent (BOLD) activity in the ventral striatum during task events was correlated with subsequent happiness ratings (Figure 1B). This brain area is a major target for dopamine neurons known to represent RPEs (Hart, Rutledge, Glimcher, & Phillips, 2014; Schultz, Dayan, & Montague, 1997) and dopamine activity is linked to BOLD activity (Knutson & Gibbs, 2007) also known to represent RPEs (Caplin, Dean, Glimcher, & Rutledge, 2010; Rutledge, Dean, Caplin, & Glimcher, 2010). Optogenetic stimulation of dopamine neurons leads to BOLD activity in the striatum and this activity is eliminated if D1 and D2 receptors are blocked (Ferenczi et al., 2016). A pharmacological study using a similar paradigm demonstrated that boosting dopamine levels increased the happiness that resulted from small rewards, consistent with the possibility that dopamine contributes to the link between rewards and mood (Rutledge, Skandali, Dayan, & Dolan, 2015).

Mood has been shown to bias perception of both potential and experienced rewards. When subjects are in a good mood, the impact of rewards on future choices is greater than when subjects are in a bad mood (Eldar & Niv, 2015). Subjects played two sets of slot machines with similar reward probabilities before and after a wheel-of-fortune draw for $7. Subjects who won the draw, and for whom mood improved, preferred the second set of slot machines, consistent with rewards from those machines being perceived as better when subjects were in a good mood. Subjects who lost the draw preferred the first set of slot machines, the ones they played before the draw. Furthermore, BOLD responses to rewards from the second set of slot machines were higher in subjects in a good mood compared to subjects in a bad mood. These results are consistent with a link between depression and a reduced impact of rewards on subsequent behavior (Dombrovski et al., 2013; Vrieze et al., 2013). Positive mood increased risk taking in the lab (Isen & Patrick, 1983) and unexpected positive outcomes outside of the laboratory (unexpected sports team wins and unexpected sunny days) increased real-world gambling (Otto, Fleming, & Glimcher, 2016). These results suggest that changes in mood resulting from one stimulus can influence judgements about potentially unrelated stimuli.

The adaptive function of mood remains unknown but one theory is that mood represents the momentum of reward, and this quantity (whether an environment is getting better or worse) can improve learning in changing environments (Eldar, Rutledge, Dolan, & Niv, 2016). Reinforcement learning algorithms allow agents to learn about the environment and to determine which actions are most likely to lead to reward. However, an inefficiency may arise when there are correlations between rewards in different states, as is often the case in the real world. For example, rain may cause fruit to grow on all trees at the same time, and snow may reduce the availability of food everywhere. A foraging animal that finds several unexpected fruits in one tree will have an improved mood as a result of these positive RPEs and will quickly update value estimates for nearby trees after also finding fruits there. The expectations that animals have then reflects not only their recent history with a particular tree, but recent changes in the overall availability of reward in the environment. Negative momentum in the environment and the resulting low mood would lead to rewards being perceived as worse than they are, allowing expectations to quickly be adjusted downward when the environment gets worse (e.g., winter is coming).

According to this theory, mood increases the efficiency of learning if the duration and intensity of moods are appropriate. Even after major life events (e.g., winning the lottery), expectations should be updated so that surprises do not continue indefinitely. Happiness does eventually return to baseline levels after major life events (Brickman, Coates, & Janoff-Bulman, 1978; Lykken & Tellegen, 1996). Problems can arise when mood persists. If a positive mood persists, rewards are perceived as better than they are, and positive-feedback dynamics lead to overly high expectations. When mood eventually stabilizes, optimistic expectations lead to a high number of negative surprises, resulting in a low mood and further positive-feedback dynamics that lead to pessimistic expectations. Such mood cycles are present in individuals suffering from bipolar disorder even in a stable environment. Another theory is that mood plays an important role in the pursuit of goals by reflecting error signals in progress toward goals and away from threats (Carver, 2015). According to this theory, a low mood should lead to increased effort and a high mood should lead to decreased effort and switching to new goals. Consistent with this theory, healthy individuals but not individuals with bipolar disorder do decrease effort after unexpected positive progress toward goals (Fulford, Johnson, Llabre, & Carver, 2010). Although theories about the function of mood are based largely on work in healthy individuals, they may allow predictions to be made about the dynamics of mood, learning, and behavior in psychiatric disorders which can be tested by researchers in computational psychiatry. Computational models of subjective feelings may also be useful for understanding a variety of psychiatric disorders. For example, one possibility is that anxiety reflects the momentum of aversive outcomes, but it is unknown how feelings of anxiety quantitatively relate to an individual's history of aversive events or to subsequent behaviour (although see (Browning, Behrens, Jocham, O'Reilly, & Bishop, 2015)).

## Bayesian inference and hierarchical models

The brain's major task is to infer the state of the world and to use that inference to make decisions. However, neither the brain's sensory data nor its prior knowledge is completely reliable. The optimal combination of these different sources of uncertain information is given by Bayes' theorem, in which a 'prior' (the initial expectation of the state of the environment) is combined with a 'likelihood' (the probability of the sensory input, given that expectation), each weighted by its precision (inverse variance), to compute a 'posterior' (an updated estimation of the state of the environment).

The brain's prior beliefs can respect the hierarchical structure of the world and of its sensory data if they take the form of a hierarchical model. Hierarchical generative models can use predictive coding to infer the causes of low-level sensory data by exploiting their high-level descriptions (Friston, 2008; Rao & Ballard, 1999). In predictive coding, a unit at a given hierarchical level sends messages to units at lower levels which predict the activity of those lower units; discrepancies between these predictions and the actual input to those lower units are then passed back up the hierarchy in the form of prediction errors. Prediction errors revise the higher-level predictions, and this hierarchical message passing continues in an iterative fashion.

Exactly which predictions ought to be changed in order to explain away a given prediction error is a crucial question for hierarchical models. An approximately Bayesian solution to this problem is to make the biggest updates to the level in the hierarchy whose uncertainty is greatest relative to the uncertainty of incoming data at the level below (i.e. if you are very uncertain about your beliefs, but your source is very reliable, you should change your beliefs a lot) (Mathys, Daunizeau, Friston, & Stephan, 2011).

A classic experiment illustrates hierarchical inference. Imagine you are shown two jars of beads: one contains 85% green and 15% red beads, the other 85% red and 15% green. The jars are hidden and a sequence of beads is drawn with replacement: GGRGGRRRRRGGGRGGGGGGRGGGG. From this sequence it appears that the jar being drawn from changes from one that is predominantly green (5 draws), to predominantly red (5 draws), to predominantly green (remaining sequence). Now suppose that although the real proportions are 85% and 15%, a malicious experimenter did not show you the jars and misleadingly told you that the proportions are 99% and 1%. You might reasonably conclude that the jar the beads were being drawn from had actually changed eight times – whenever the color changed.

This is what happens when the precision at the bottom of a hierarchical model is too high relative to the precision at the top. Following a sensory prediction error, the model concludes there must have been a change in the environment, rather than ‘putting it down to chance’. This precision imbalance might contribute to various phenomena observed in schizophrenia.

## Schizophrenia, precision, and inference

We now explore how neurobiological abnormalities in schizophrenia might be characterized in computational terms, and how these characterisations might help us understand the disorder (Adams et al., 2015). We discuss reductions in synaptic gain in higher areas in the hierarchy, the notion of aberrant salience, and probabilistic inference. Other excellent reviews explore these subjects in more detail (e.g., reinforcement learning models and schizophrenia (Corlett & Fletcher, 2014; Deserno, Boehme, Heinz, & Schlagenhauf, 2013), models of negative symptoms (e.g. apathy) (Strauss, Waltz, & Gold, 2014), and biophysical models (Cohen, Braver, & O’Reilly, 1996; Wang & Krystal, 2014)).

What are the main cortical abnormalities in schizophrenia and what do they have in common (Adams, Stephan, Brown, Frith, & Friston, 2013)? One key abnormality is thought to be hypofunction of the N-methyl-D-aspartate receptor (NMDA-R), a glutamate receptor with profound effects on both synaptic gain (due to its prolonged opening time) and synaptic plasticity (via long term potentiation or depression) in both the prefrontal cortex and hippocampus. Synaptic gain (or short-term synaptic plasticity (Stephan, Baldeweg, & Friston, 2006)) refers to a multiplicative change in the influence of presynaptic input on postsynaptic responses. A second abnormality is the reduced synthesis of γ-aminobutyric acid (GABA) by inhibitory interneurons in prefrontal cortex. A third is the hypoactivation of $D_1$ receptors in prefrontal cortex (we shall discuss striatal hyperactivation of $D_2$ receptors in the next section).

These abnormalities could all reduce synaptic gain in prefrontal cortex or hippocampus, which are positioned near the top of the cortical hierarchy. NMDA-R hypofunction and $D_1$ receptor hypoactivity are most easily related to a change in synaptic gain. Similarly, a GABAergic deficit might cause a loss of ‘synchronous’ gain: sustained oscillations in neuronal populations are facilitated through their rhythmic inhibition by GABAergic interneurons, putatively increasing communication between neurons that oscillate in phase (Fries, 2005).

How can synaptic gain (and its loss) be understood in computational terms? One answer rests on the idea that the brain approximates and simplifies Bayesian inference by using probability distributions that can be encoded by a few ‘sufficient statistics’ (e.g., the mean and its precision). Whilst precision determines the influence one piece of information has over another in Bayesian inference, synaptic gain determines the influence one neural population has over another in neural message passing. The neurobiological substrate of precision could therefore be synaptic gain (Feldman & Friston, 2010), and a loss of synaptic gain in a given area could reduce the precision of information encoded there.

A loss of synaptic gain in prefrontal cortex or hippocampus would reduce the influence of their inputs on lower-level areas. In the brain's hierarchical model, this would correspond to a loss of influence (i.e., precision) of the model's priors over the sensory data. This simple computational change (i.e. a loss of precision of prior beliefs or a relative increase in the precision of sensory data) can describe a great variety of phenomena in schizophrenia (see Figure 2; more references and predictive coding simulations of some of these phenomena are elsewhere (Adams et al., 2013)):

- At a neurophysiological level, responses to predictable stimuli resemble responses to unpredicted stimuli, and *vice versa*, in perceptual electrophysiology experiments (e.g., the P50 or P300 responses to tones (Turetsky et al., 2007));
- At a network level, higher regions of cortex (i.e. prefrontal cortex and hippocampus) have diminished connectivity to the thalamus subjects with schizophrenia relative to control subjects, whereas primary sensory areas are coupled more strongly with the thalamus (Anticevic et al., 2014);
- At a perceptual level, a greater resistance to visual illusions (Silverstein & Keane, 2011) (which exploit the effects of visual priors on ambiguous images, for example the famous "hollow-mask" illusion (Dima et al., 2009)) and a failure to attenuate the sensory consequences of one's own actions, which could diminish one's sense of agency (Shergill, Samson, Bays, Frith, & Wolpert, 2005);
- At a behavioral level, impaired smooth visual pursuit of a predictably moving target, but improved tracking of a sudden unpredictable change in a target's motion (Hong et al., 2008).

Another way of modeling the effects of NMDA-R hypofunction is to use a biophysical model that contains specific parameters for receptor conductances using, for example, a spiking neural network model of spatial working memory (Murray et al., 2014). If NMDA-R conductance between pyramidal cells and interneurons is reduced in this model there is a loss of precision of the spatial location over time. This predicts an increase in false alarms to local distractors in a spatial working memory task, but no change in the number of missed trials. These effects are observed in healthy subjects given ketamine (an NMDA-R antagonist) and are similar to those found in schizophrenia (Mayer & Park, 2012).

How do these ideas relate to the symptoms of psychosis? A reasonable hypothesis would be that a loss of high-level precision might result in diffuse generalized cognitive problems (as routinely found in schizophrenia) and over-attention to sensory stimuli (as found in the 'delusional mood' (Corlett & Fletcher, 2014) and also in autism), and overadjustments of beliefs following chance events. However, one might expect that these belief updates should be fleeting (as they would be vulnerable to rapid updating) unlike delusions. It may therefore be that the permanence of delusions reflects other abnormal learning mechanisms in prefrontal cortex or striatum (discussed elsewhere (Adams et al., 2015)).

**Aberrant salience and psychosis.**

The best-established neurobiological abnormality in schizophrenia is an increased presynaptic availability of dopamine in the associative striatum that correlates with positive symptoms (Howes & Kapur, 2009). Kapur proposed that this hyperdopaminergia causes a state of 'aberrant salience', in which perception of external stimuli and 'internal' thoughts could take on undue significance, due to an increased stimulus-independent release of dopamine (Kapur, 2003). The stimuli or thoughts themselves could be quite innocuous (e.g., a street light turning on, or wondering if one should buy some food) but the experience of salience coincident with the stimulus would drive the subject to seek (often delusional) explanations, such as, "That light turning on means I am the Son of God". The aberrant salience hypothesis was based on the theory of incentive salience, which proposes that dopaminergic activity gives motivational impetus to act on stimuli whose values have already been learned (Berridge, 2007).

To test Kapur's hypothesis, Roiser developed the Salience Attribution Test, which assesses the extent to which subjects explicitly (consciously) learn associations between stimulus attributes and outcomes (some attributes are predictive of outcomes but some are not), and the extent to which reaction times are affected by the same attributes. The Salience Attribution Test thus assesses both explicit learning and implicit motivation. Both medicated (Roiser et al., 2009) and unmedicated (Roiser, Howes, Chaddock, Joyce,, & McGuire, 2013) prodromal schizophrenic subjects showed greater aberrant salience in the explicit rather than implicit measures and the former related to their delusions or abnormal thoughts. Although others have found aberrant motivational salience abnormalities in schizophrenia (Pankow et al., 2015), the most consistent finding in patients is a loss of adaptive motivational salience: reaction times do not decrease to rewarding stimuli, as they do in control subjects (Smieskova et al., 2015).

Overall, it seems that abnormal motivational salience may exist in schizophrenia, but it is less clear that this could be a cause of positive symptoms such as delusions and, in particular, hallucinations. Aberrant motivational salience may work best as an account of manic psychosis, in which the subject is energized and perceives events in a positive light, rather than schizophrenic psychosis, which is often aversive in nature. Conversely, diminished adaptive motivational salience provides a plausible explanation for negative symptoms.

What other kinds of salience might be abnormal in schizophrenia, and how might they be cast in computational terms? Some possibilities, all proposed to relate to striatal dopamine release (Schwartenbeck, FitzGerald, & Dolan, 2016; Winton-Brown, Fusar-Poli, Ungless, & Howes, 2014), are:

- Reward and aversive prediction error. While many dopamine neurons are phasically active in response to unexpected rewards, others are active following unexpected aversive events, and some are active after both, and so have been said to represent a salience signal. However, in the Rescorla-Wagner model of conditioning, salience refers not to the absolute value of the error $\left|(r-V)\right|$ but to the associability $\alpha_{CS_i}$ of the conditioned stimulus property with the value update (with learning rate $\beta$): $\Delta V_{CS_i} = \alpha_{CS_i}\beta(r-V)$. In more sophisticated models of belief updating, static $\alpha$ and $\beta$ parameters are replaced by dynamic precision estimates to give precision-weighted prediction errors (Mathys et al., 2011). Thus aberrant salience may be cast as aberrant precision-weighting.

- Surprise or novelty signals. Surprise and novelty are often conflated but are fundamentally different. They are computed by comparing the current event with one's expectation or memory, respectively (Barto, Mirolli, & Baldassarre, 2013). In information theory, 'surprisal' is the negative log probability of an event. For observation $o$ and model $m$ with parameters $\theta$ this corresponds to . Aberrant novelty or surprisal signaling could lead to attentional problems but a direct link to delusions is not intuitive, as aberrantly novel or surprising events need not cause aberrant learning.

- Informational salience. Bayesian surprise is formalized as the information used to transform a prior into a posterior distribution (Itti & Baldi, 2009); i.e. the Kullback-Leibler divergence between these distributions following an observation: $KL[p(\theta \mid o,m) \,||\, p(\theta \mid m)]$. Informational salience, or the 'epistemic value' of observations (Friston et al., 2015), is a promising candidate for the kind of salience one might expect to be aberrant in schizophrenia, as it heralds not just surprising observations, but also shifts in beliefs (Schwartenbeck et al., 2016).

Of these accounts, only reinforcement learning has been explored in any depth in schizophrenia (Deserno et al., 2013). One well-replicated abnormality is a reduction in neural activity in the ventral striatum during RPE signaling and reward anticipation in schizophrenia that correlates with negative symptoms (Juckel et al., 2006). However, despite abnormal neuroimaging results, behavior is often similar to that of controls (Murray, Corlett, & Fletcher, 2010). Indeed, reinforcement learning models serve best as explanations of negative symptoms (Strauss et al., 2014), including pronounced asymmetry in learning (that is, a failure to learn stimulus-reward associations but intact learning of stimulus-punishment associations), failure to infer the values of actions (cf. anhedonia in depression), greater discounting of rewards that require effort (Hartmann et al., 2015), and a loss of uncertainty-driven exploration such that valuable states are never discovered. One important caveat here is the recent finding that when working memory is incorporated into a reinforcement learning model, patients with schizophrenia only showed deficits in working memory parameters and not parameters related to reinforcement learning (Collins, Brown, Gold, Waltz, & Frank, 2014).

### Probabilistic reasoning in schizophrenia.

The most widely used test of probabilistic inference in schizophrenia is the beads task, in which the subject has to guess which of two jars a sequence of red and green beads is coming from: the 'red' jar containing mostly red and some green beads, or the 'green' jar containing the reverse distribution. The task has two popular variants: in the 'draws to decision' version, the subject stops the sequence when sure of the color of the jar, and the number of beads seen is recorded. In the 'probability estimates' version, the subject sees a long sequence of beads drawn from one jar (with or without a change of jar at some point) and has to estimate the probability of either jar being the source after every bead.

In a well-replicated finding, around half of subjects with schizophrenia decide on the jar color (in the draws to decision task) after seeing only one or two beads, a so-called jumping-to-conclusions bias (Fine, Gardner, Craigie, & Gold, 2007). This has been interpreted as the effect of overweighting evidence, i.e. making larger updates to beliefs than are warranted by the data. This explanation is unlikely to be the whole reason for the effect, however, as any increased belief updating in the probability estimates version is of relatively small magnitude (Fine et al., 2007), and in a related paradigm, schizophrenic subjects learn *less* from positive feedback than controls, *especially* those subjects that jump to conclusions (Averbeck, Evans, Chouhan, Bristow, & Shergill, 2010).

Other possible but underexplored explanations for the jumping-to-conclusions bias are that schizophrenic subjects care less about making an incorrect decision, or that they are more reluctant to ask for another bead, or have greater stochasticity in their decision process. A computational model that incorporated the costs of sampling and wrong decisions and included a stochasticity (softmax) parameter $\tau$ revealed that only a higher $\tau$ accounted for the early decisions made by schizophrenic subjects in the draws to decision task (Moutoussis, Bentall, El-Deredy, & Dayan, 2011). This could reflect a loss of synaptic gain (i.e., precision encoding) in prefrontal cortex or the striatum (FitzGerald, Schwartenbeck, Moutoussis, Dolan, & Friston, 2015).

Overall, while much progress has been made, we are far from a complete computational account of delusions. Many delusions arise too quickly to be explained using incremental belief updating, and others seem to come from memory. It is also unclear how delusions become so persistent (Adams et al., 2015). Furthermore, computational accounts will not be complete until they also cover their distinctive themes (e.g., persecution, grandiosity, self-reference, etc.).

### Computational phenotyping using social games

An important aspect of human cognition is the ability to model and understand the behavior of other humans (Saxe & Kanwisher, 2003). This ability plays an important role in both cooperation and competition and may be affected in a range of psychiatric disorders including autism spectrum disorder (ASD), borderline personality disorder (BPD), schizophrenia, and depression. Social games with multiple interacting agents permit computational modeling and neuroimaging of inter-personal exchange to characterize social interactions. Using computational models, behavioral and neural parameters can be estimated that might allow computational phenotyping (Montague et al., 2012), whereby subjects are categorized according to inter-individual differences in the computational mechanisms that underlie, for example, their social interactions.

One of the most popular tasks for studying the neural mechanisms that underlie human interactions is the trust game (Ruff & Fehr, 2014). In a typical experiment, subjects play ten rounds of the trust game with the same partner. In each round, the investor is endowed with some money and decides how much of that endowment to entrust to the trustee. Any money received by the trustee from the investor is tripled, and the trustee then decides how much to return to the investor. In this way, cooperation is desirable for both players, and investors should invest if they can expect trustees to return money. Players able to make inferences about the likely mental state of their partner should have an advantage in maximizing earnings. When subjects that suffer from BPD play the multi-round trust game with healthy participants, BPD subjects are unable to maintain cooperation (King-Casas et al., 2008). Neural activity in the anterior insula, a region often found to respond to norm violations (Xiang, Lohrenz, & Montague, 2013), differed between BPD and healthy players. In healthy trustees, insula activity was highest when investors made small offers. Insula activity in BPD trustees did not reflect the size of investor offers, and this dysfunction in neural responses to partner decisions may explain the difficulty of BPD subjects to develop trust with social partners. In ASD trustees playing a trust game with healthy investors, neural activity in the cingulate cortex was lower than in healthy trustees both when investor decisions were revealed and when trustee repayment decisions were made (Chiu et al., 2008), and this signal has been linked to the ability to model one's own social intentions, an important capacity for social interaction.

Computational models of social behavior provide trial-by-trial predictions of task variables that depend on individual model parameters. These predictions can be used to probe the neural mechanisms that underlie differences present in psychiatric disorders. For example, trust game decisions can be fitted to a computational model that produces an estimate of a player's depth-of-thought, which captures the richness of the models a player builds about a partner (Xiang, Ray, Lohrenz, Dayan, & Montague, 2012). A level 0 subject does not simulate partner choices. A level 1 subject assumes a level 0 partner and simulates partner choices accordingly. A level 2 subject assumes a level 1 partner and simulates accordingly. Applying this model to data from healthy investors playing games anonymously with either healthy trustees or BPD trustees revealed that the depth-of-thought of healthy investors was level 0 less than 20% of the time when playing with healthy trustees but level 0 more than 60% of the time when playing with BPD trustees. This finding indicates that although healthy subjects are capable of simulating partner choices, cooperation breaks down when playing with BPD trustees as revealed by computational models. Healthy investor behavior distinguishes not only between BPD and healthy trustees but also between anonymous trustees with a variety of psychiatric diagnoses including BPD, ASD, attention deficit hyperactivity disorder, and depression, leading to the suggestion that healthy individuals might act as a sort of "biosensor" in social games and can help to identify the important differences between patient groups (Koshelev, Lohrenz, Vannucci, & Montague, 2010).

### Summary

Computational psychiatry aims to develop mathematical models that are useful for understanding psychiatric disorders and for bridging the gap between clinical practice and basic neuroscience research. New models will capture dysfunction in a way that allows current clinical definitions of psychiatric disorders to be updated with definitions that map more closely to the neural circuits that perform the aberrant computations. Much research focuses on the aberrant decision making present in many psychiatric disorders. We have described several different computational approaches to the study of psychiatric disorders, providing examples of research related to features of depression, schizophrenia, borderline personality disorder, and autism spectrum disorder. The global burden of mental illness is expected to increase, but few novel treatments for psychiatric disorders have been introduced in recent years. The hope of computational psychiatry is that models will link specific behaviors to specific activity in neural circuits, providing new insights that facilitate the development of new treatments for psychiatric disorders.

## References

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2015). Computational psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery, and Psychiatry*, *87*, 53-63. doi:10.1136/jnnp-2015-310737

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

Anticevic, A., Cole, M. W., Repovs, G., Murray, J. D., Brumbaugh, M. S., Winkler, A. M., … Glahn, D. C. (2014). Characterizing thalamo-cortical disturbances in schizophrenia and bipolar illness. *Cerebral Cortex*, *24*, 3116-3130. doi:10.1093/cercor/bht165

Averbeck, B. B., Evans, S., Chouhan, V., Bristow, E., & Shergill, S. S. (2010). Probabilistic learning and inference in schizophrenia. *Schizophrenia Research*, *127*, 115-122. doi:10.1016/j.schres.2010.08.009

Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or surprise? *Frontiers in Psychology*, *4*, 907. doi:10.3389/fpsyg.2013.00907

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology*, *191*, 391-431. doi:10.1007/s00213-006-0578-x

Brickman, P., Coates, D., & Janoff-Bulman, R. (1978). Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology*, *36*, 917-927. doi:10.1037/0022-3514.36.8.917

Brown, H. R., Zeidman, P., Smittenaar, P., Adams, R. A., McNab, F., Rutledge, R. B., & Dolan, R. J. (2014). Crowdsourcing for cognitive science - The utility of smartphones. *PLoS ONE*, *9*, e100662. doi:10.1371/journal.pone.0100662

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, *18*, 590-596. doi:10.1038/nn.3961

Caplin, A., Dean, M., Glimcher, P. W., & Rutledge, R. B. (2010). Measuring beliefs and rewards: A neuroeconomic approach. *The Quarterly Journal of Economics*, *125*, 923-960. doi:10.1162/qjec.2010.125.3.923

Carver, C. S. (2015). Control processes, priority management, and affective dynamics. *Emotion Review*, *7*, 301-307. doi:10.1177/1754073915590616

Chiu, P. H., Kayali, M. A., Kishida, K. T., Tomlin, D., Klinger, L. G., Klinger, M. R., & Montague, P. R. (2008). Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism. *Neuron*, *57*, 463-473. doi:10.1016/j.neuron.2007.12.020

Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control and schizophrenia: Recent developments and current challenges. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *351*, 1515-1527. doi:10.1098/rstb.1996.0138

Cohen, J. Y., Amoroso, M. W., & Uchida, N. (2015). Serotonergic neurons signal reward and punishment on multiple timescales. *eLife*, *4*, e06346. doi:10.7554/eLife.06346

Collins, A. G. E., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory: Contributions to reinforcement learning impairments in schizophrenia. *The Journal of Neuroscience: 34*, 13747-13756. doi:10.1523/JNEUROSCI.0989-14.2014

Corlett, P. R., & Fletcher, P. C. (2014). Computational psychiatry: A Rosetta Stone linking the brain to mental illness. *The Lancet Psychiatry*. doi:10.1016/S2215-0366(14)70298-6

Cowen, P. J., & Browning, M. (2015). What has serotonin to do with depression? *World Psychiatry*, *14*, 158-160. doi:10.1002/wps.20229

Crockett, M., & Cools, R. (2015). Serotonin and aversive processing in affective and social decision-making. *Current Opinion in Behavioral Sciences*, *5*, 64-70. doi:10.1016/j.cobeha.2015.08.005

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*, *175*, 526-536.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*, 126. doi:10.1186/1741-7015-11-126

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*, 1204-1215. doi:10.1016/j.neuron.2011.02.027

Dayan, P., Dolan, R. J., Friston, K. J., & Montague, P. R. (2015). Taming the shrewdness of neural function: methodological challenges in computational psychiatry. *Current Opinion in Behavioral Sciences*, *5*, 128–132. doi:10.1016/j.cobeha.2015.09.009

Deserno, L., Boehme, R., Heinz, A., & Schlagenhauf, F. (2013). Reinforcement learning and dopamine in schizophrenia: dimensions of symptoms or specific features of a disease group? *Frontiers in Psychiatry*, *4*, 172. doi:10.3389/fpsyt.2013.00172

Dima, D., Roiser, J. P., Dietrich, D. E., Bonnemann, C., Lanfermann, H., Emrich, H. M., & Dillo, W. (2009). Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. *NeuroImage*, *46*, 1180-1186. doi:10.1016/j.neuroimage.2009.03.033

Dombrovski, A. Y., Szanto, K., Clark, L., Reynolds, C. F., & Siegle, G. J. (2013). Reward signals, attempted suicide, and impulsivity in late-life depression. *JAMA Psychiatry*, *70*, 1020-1030. doi:10.1001/jamapsychiatry.2013.75

Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications*, *6*, 6149. doi:10.1038/ncomms7149

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences*, *20*, 15–24. doi:10.1016/j.tics.2015.07.010

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215. doi:10.3389/fnhum.2010.00215

Ferenczi, E. A., Zalocusky, K. A., Liston, C., Grosenick, L., Warden, M. R., Amatya, D., … Deisseroth, K. (2016). Prefrontal cortical regulation of brainwide circuit dynamics and reward-related behavior. *Science*, *351*, aac9698. doi:10.1126/science.aac9698

Fine, C., Gardner, M., Craigie, J., & Gold, I. (2007). Hopping, skipping or jumping to conclusions? Clarifying the role of the JTC bias in delusions. *Cognitive Neuropsychiatry*, *12*, 46-77. doi:10.1080/13546800600750597

FitzGerald, T. H. B., Schwartenbeck, P., Moutoussis, M., Dolan, R. J., & Friston, K. (2015). Active inference, evidence accumulation, and the urn task. *Neural Computation*, *27*, 306–328. doi:10.1162/NECO_a_00699

Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., … Yager, J. (2013). The initial field trials of DSM-5: New blooms and old thorns. *American Journal of Psychiatry*, *170*(1), 1–5. doi:10.1176/appi.ajp.2012.12091189

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, *9*, 474-480. doi:10.1016/j.tics.2005.08.011

Friston, K. J. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*, e1000211. doi:10.1371/journal.pcbi.1000211

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*, 187-214. doi:10.1080/17588928.2015.1020053

Fulford, D., Johnson, S. L., Llabre, M. M., & Carver, C. S. (2010). Pushing and coasting in dynamic goal pursuit coasting is attenuated in bipolar disorder. *Psychological Science*, *21*, 1021-1027. doi:10.1177/0956797610373372

Gradin, V. B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., … Steele, J. D. (2011). Expected value and prediction error abnormalities in depression and schizophrenia. *Brain*, *134*, 1751-1764. doi:10.1093/brain/awr059

Harmer, C. J., O'Sullivan, U., Favaron, E., Rachel Massey-Chase, B. A., Ayres, R., Reinecke, A., … others. (2009). Effect of acute antidepressant administration on negative affective bias in depressed patients. *The American Journal of Psychiatry*, *166*(10), 1178–1184. doi:10.1176/appi.ajp.2009.09020149

Hart, A. S., Rutledge, R. B., Glimcher, P. W., & Phillips, P. E. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *The Journal of Neuroscience*, *34*, 698-704. doi:10.1523/JNEUROSCI.2489-13.2014

Hartmann, M. N., Hager, O. M., Reimann, A. V., Chumbley, J. R., Kirschner, M., Seifritz, E., … Kaiser, S. (2015). Apathy but not diminished expression in schizophrenia is associated with discounting of monetary rewards by physical effort. *Schizophrenia Bulletin*, *41*(2), 503–512. doi:10.1093/schbul/sbu102

Hong, L. E., Turano, K. A., O'Neill, H., Hao, L., Wonodi, I., McMahon, R. P., … Thaker, G. K. (2008). Refining the predictive pursuit endophenotype in schizophrenia. *Biological Psychiatry*, *63*, 458-464. doi:10.1016/j.biopsych.2007.06.004

Howes, O. D., & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: version III--the final common pathway. *Schizophrenia Bulletin*, *35*, 549-562. doi:10.1093/schbul/sbp006

Huys, Q. J. M., Guitart-Masip, M., Dolan, R. J., & Dayan, P. (2015). Decision-theoretic psychiatry. *Clinical Psychological Science*, *3*, 400-421. doi:10.1177/2167702614562040

Huys, Q. J., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? *Neural Networks*, *24*, 544-551. doi:10.1016/j.neunet.2011.03.001

Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: A behavioural meta-analysis. *Biology of Mood and Anxiety Disorders*, *3*, 12. http:/doi.org/10.1186/2045-5380-3-12

Isen, A. M., & Patrick, R. (1983). The effect of positive feelings on risk taking: When the chips are down. *Organizational Behavior and Human Performance*, *31*, 194-202. doi:10.1016/0030-5073(83)90120-4

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*, 1295-1306. doi:10.1016/j.visres.2008.09.007

Juckel, G., Schlagenhauf, F., Koslowski, M., Wüstenberg, T., Villringer, A., Knutson, B., … Heinz, A. (2006). Dysfunction of ventral striatal reward prediction in schizophrenia. *NeuroImage*, *29*, 409-416. doi:10.1016/j.neuroimage.2005.07.051

Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *The American Journal of Psychiatry*, *160*, 13-23. doi:10.1176/appi.ajp.160.1.13

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*, 932-932. doi:10.1126/science.1192439

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, *321*, 806-810. doi:10.1126/science.1156902

Knutson, B., & Gibbs, S. E. B. (2007). Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacology*, *191*, 813-822. doi:10.1007/s00213-006-0686-7

Koshelev, M., Lohrenz, T., Vannucci, M., & Montague, P. R. (2010). Biosensor approach to psychopathology classification. *PLoS Computational Biology*, *6*, e1000966. doi:10.1371/journal.pcbi.1000966

Lykken, D., & Tellegen, A. (1996). Happiness is a stochastic phenomenon. *Psychological Science*, *7*, 186-189. doi:10.1111/j.1467-9280.1996.tb00355.x

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*, 154-162. doi:10.1038/nn.2723

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39. doi:10.3389/fnhum.2011.00039

Mayer, J. S., & Park, S. (2012). Working memory encoding and false memory in schizophrenia and bipolar disorder in a spatial delayed response task. *Journal of Abnormal Psychology*, *121*, 784-794. doi:10.1037/a0028836

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*, 72-80. doi:10.1016/j.tics.2011.11.018

Moutoussis, M., Bentall, R. P., El-Deredy, W., & Dayan, P. (2011). Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. *Cognitive Neuropsychiatry*, *16*, 422-447. doi:10.1080/13546805.2010.548678

Murray, C. J. L., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., … Lopez, A. D. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, *380*, 2197-2223. doi:10.1016/S0140-6736(12)61689-4

Murray, G. K., Corlett, P. R., & Fletcher, P. C. (2010). The neural underpinnings of associative learning in health and psychosis: how can performance be preserved when brain responses are abnormal? *Schizophrenia Bulletin*, *36*, 465-471. doi:10.1093/schbul/sbq005

Murray, J. D., Anticevic, A., Gancsos, M., Ichinose, M., Corlett, P. R., Krystal, J. H., & Wang, X.-J. (2014). Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. *Cerebral Cortex*, *24*, 859-872. doi:10.1093/cercor/bhs370

Otto, A. R., Fleming, S. M., & Glimcher, P. W. (2016). Unexpected but incidental positive outcomes predict real-world gambling. *Psychological Science*, *27*, 299-311. doi:10.1177/0956797615618366

Pankow, A., Katthagen, T., Diner, S., Deserno, L., Boehme, R., Kathmann, N., … Schlagenhauf, F. (2015). Aberrant salience is related to dysfunctional self-referential processing in psychosis. *Schizophrenia Bulletin*, *42*, 67-76. doi:10.1093/schbul/sbv098

Pareés, I., Brown, H., Nuruki, A., Adams, R. A., Davare, M., Bhatia, K. P., … Edwards, M. J. (2014). Loss of sensory attenuation in patients with functional (psychogenic) movement disorders. *Brain*, *137*, 2916-2921. doi:10.1093/brain/awu237

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*, 545-556. doi:10.1038/nrn2357

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

Roiser, J. P., Howes, O. D., Chaddock, C. A., Joyce, E. M., & McGuire, P. (2013). Neural and behavioral correlates of aberrant salience in individuals at risk for psychosis. *Schizophrenia Bulletin*, *39*, 1328-1336. doi:10.1093/schbul/sbs147

Roiser, J. P., Stephan, K. E., den Ouden, H. E. M., Barnes, T. R. E., Friston, K. J., & Joyce, E. M. (2009). Do patients with schizophrenia exhibit aberrant salience? *Psychological Medicine*, *39*, 199-209. doi:10.1017/S0033291708003863

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*, 549-562. doi:10.1038/nrn3776

Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *The Journal of Neuroscience*, *30*, 13525-13536.

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences USA*, *111*, 12252-12257. doi:10.1073/pnas.1407535111

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2015). Dopaminergic modulation of decision making and subjective well-being. *The Journal of Neuroscience*, *35*, 9811-9822. doi:10.1523/JNEUROSCI.0702-15.2015

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*, 1835-1842. doi:10.1016/S1053-8119(03)00230-1

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593-1599. doi:10.1126/science.275.5306.1593

Schwartenbeck, P., FitzGerald, T. H. B., & Dolan, R. (2016). Neural signals encoding shifts in beliefs. *NeuroImage*, *125*, 578-586. doi:10.1016/j.neuroimage.2015.10.067

Shergill, S. S., Samson, G., Bays, P. M., Frith, C. D., & Wolpert, D. M. (2005). Evidence for sensory prediction deficits in schizophrenia. *The American Journal of Psychiatry*, *162*, 2384-2386. doi:10.1176/appi.ajp.162.12.2384

Silverstein, S. M., & Keane, B. P. (2011). Perceptual organization impairment in schizophrenia and associated brain mechanisms: review of research from 2005 to 2010. *Schizophrenia Bulletin*, *37*, 690-699. doi:10.1093/schbul/sbr052

Simon, G. E. (2003). Social and economic burden of mood disorders. *Biological Psychiatry*, *54*, 208-215. doi:10.1016/S0006-3223(03)00420-7

Smieskova, R., Roiser, J. P., Chaddock, C. A., Schmidt, A., Harrisberger, F., Bendfeldt, K., … Borgwardt, S. (2015). Modulation of motivational salience processing during the early stages of psychosis. *Schizophrenia Research*, *166*, 17-23. doi:10.1016/j.schres.2015.04.036

Stephan, K. E., Baldeweg, T., & Friston, K. J. (2006). Synaptic plasticity and dysconnection in schizophrenia. *Biological Psychiatry*, *59*, 929-939. doi:10.1016/j.biopsych.2005.10.005

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, *25*, 85-92. doi:10.1016/j.conb.2013.12.007

Strauss, G. P., Waltz, J. A., & Gold, J. M. (2014). A review of reward processing and motivational impairment in schizophrenia. *Schizophrenia Bulletin*, *40 Suppl 2*, S107–S116. doi:10.1093/schbul/sbt197

Turetsky, B. I., Calkins, M. E., Light, G. A., Olincy, A., Radant, A. D., & Swerdlow, N. R. (2007). Neurophysiological endophenotypes of schizophrenia: the viability of selected candidate measures. *Schizophrenia Bulletin*, *33*, 69-94. doi:10.1093/schbul/sbl060

Vrieze, E., Pizzagalli, D. A., Demyttenaere, K., Hompes, T., Sienaert, P., de Boer, P., … Claes, S. (2013). Reduced reward learning predicts outcome in major depressive disorder. *Biological Psychiatry*, *73*, 639-645. doi:10.1016/j.biopsych.2012.10.014

Wang, X.-J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*, 638-654. doi:10.1016/j.neuron.2014.10.018

Wiecki, T. V., Poland, J., & Frank, M. J. (2015). Model-based cognitive neuroscience approaches to computational psychiatry: clustering and classification. *Clinical Psychological Science*, *3*, 378-399. doi:10.1177/2167702614565359

Winton-Brown, T. T., Fusar-Poli, P., Ungless, M. A., & Howes, O. D. (2014). Dopaminergic basis of salience dysregulation in psychosis. *Trends in Neurosciences*, *37*, 85-94. doi:10.1016/j.tins.2013.11.003

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *The Journal of Neuroscience*, *33*, 1099-1108. doi:10.1523/JNEUROSCI.1642-12.2013

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Computational Biology*, *8*, e1002841. doi:10.1371/journal.pcbi.1002841

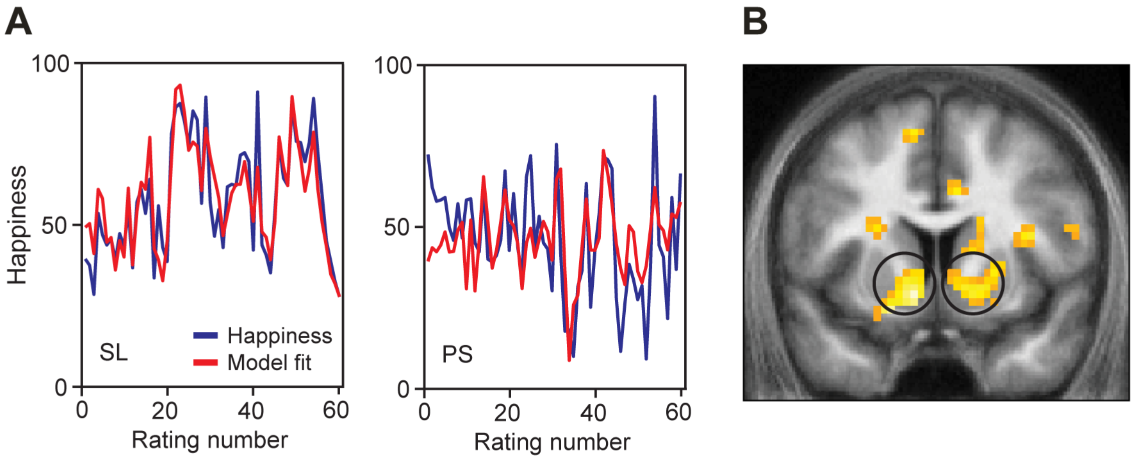

**Figure 1: Computational modeling of mood and its relation to neural activity. A)** Subjects participating in a probabilistic reward task answered the question ‘How happy are you at this moment?’ after every 2-3 trials. Happiness (as a proxy for mood) for two example subjects fluctuated over the course of the experiment. A computational model of mood dynamics accounted for subjective ratings using the recent history of rewards and expectations. Happiness depends not on how well subjects were doing in the task, but in whether they were doing better than expected. **B)** Blood oxygen levels measured with functional MRI revealed that activity during task events was correlated with subsequent happiness ratings, consistent with the possibility that neural activity in this region, the ventral striatum, relates to changes in mood. Because activity in this area is linked to dopamine, one possibility, supported by pharmacological research, is that dopamine plays a role in determining mood. Figure adapted from (Rutledge et al., 2014).

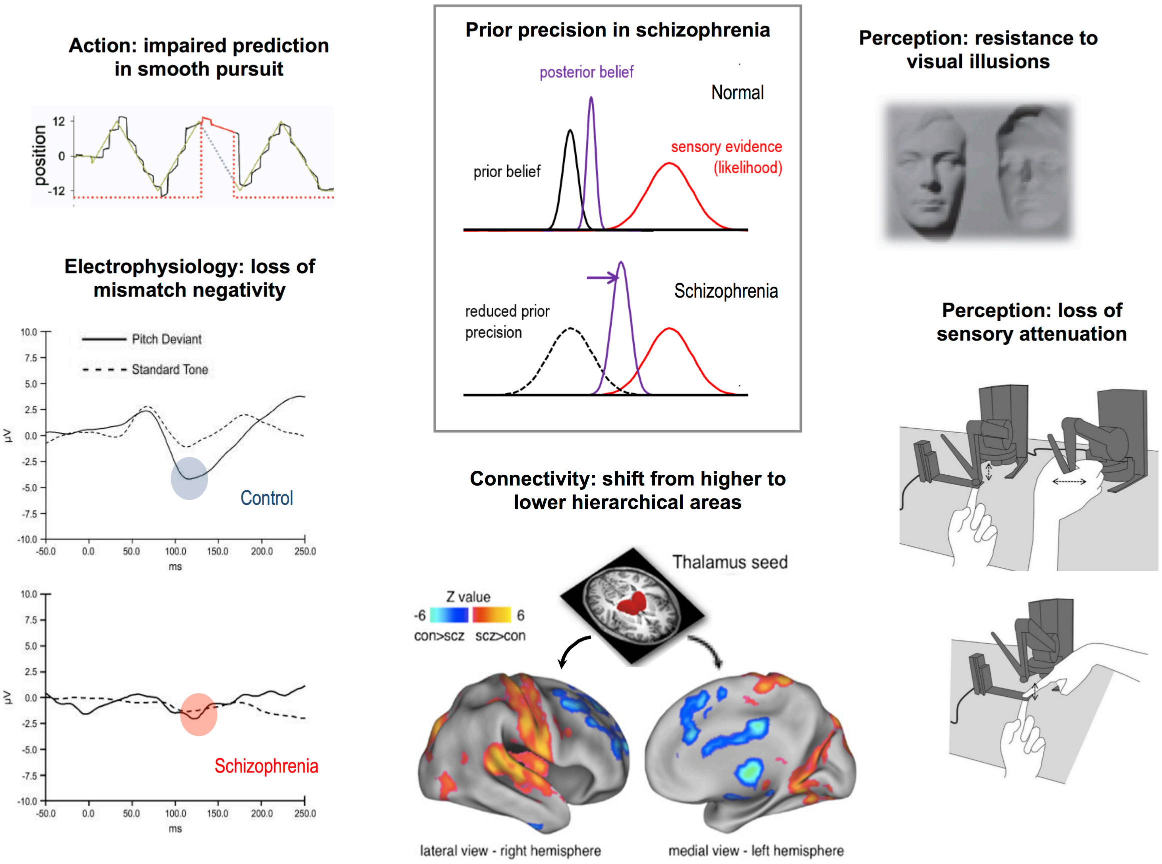

**Figure 2: Effects of a hierarchical precision imbalance in schizophrenia.** A loss of precision encoding in higher hierarchical areas would bias inference away from prior beliefs and toward sensory evidence, illustrated schematically in the middle panel. This single change could manifest in many ways (moving anticlockwise from left to right). **A)** A loss of the ability to smoothly pursue a target moving predictably. The eye position of this subject with schizophrenia frequently falls behind the target and has to saccade to catch up again. When the target is briefly stabilized on the retina (to reveal the purely predictive element of pursuit), shown as the red unbroken line, eye velocity drops significantly (figure adapted from (Hong et al., 2008)). **B)** These graphs illustrate average electrophysiological responses in a mismatch negativity paradigm, in which a deviant oddball tone follows a series of identical tones. In the control subject, the oddball causes a pronounced negative deflection at around 120 ms (blue circle), but in a subject with schizophrenia, there is no such deflection (red circle) and neural responses to predictable and unpredictable stimuli are similar (figure adapted from (Turetsky et al., 2007)). **C)** The physiological change underlying the precision imbalance is a relative decrease in synaptic gain in higher hierarchical areas, and a relative increase in lower hierarchical areas. This change would also manifest as an alteration in connectivity, shown here as whole brain differences in connectivity with a thalamic seed between controls and subjects with schizophrenia. Red/yellow areas are more strongly coupled in subjects with schizophrenia, and include sensory areas (auditory, visual, motor, and somatosensory). Blue areas are more weakly coupled, and include higher hierarchical areas (medial and lateral prefrontal cortex, cingulate cortex, and hippocampus) and the striatum (figure adapted from (Anticevic et al., 2014)). **D)** An imbalance in hierarchical precision may lead to a failure to attenuate the sensory consequences of one's own actions (Shergill et al., 2005), here illustrated with the force-matching paradigm used to measure this effect. The subject must match a target force by either pressing on a bar with their finger (below) or using a mechanical transducer (top). Control subjects exert more force than necessary in the former condition, but schizophrenic subjects do not (figure adapted from (Pareés et al., 2014)). **E)** A loss of the precision of prior beliefs can cause a resistance to visual illusions that rely on prior beliefs for their perceptual effects. Control subjects perceive the face on the right as a convex face lit from

below, due to a powerful prior belief that faces are convex, whereas subjects with schizophrenia tend to perceive the image veridically as a concave hollow face lit from above. Figure reproduced from (Adams et al., 2015).