Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

State representation in mental illness

Angela Radulescu[1] and Yael Niv[1,2]

Reinforcement learning theory provides a powerful set of computational ideas for modeling human learning and decision making. Reinforcement learning algorithms rely on state representations that enable efficient behavior by focusing only on aspects relevant to the task at hand. Forming such representations often requires selective attention to the sensory environment, and recalling memories of relevant past experiences. A striking range of psychiatric disorders, including bipolar disorder and schizophrenia, involve changes in these cognitive processes. We review and discuss evidence that these changes can be cast as altered state representation, with the goal of providing a useful transdiagnostic dimension along which mental disorders can be understood and compared.

Addresses
[1] Psychology Department, Princeton University, United States
[2] Princeton Neuroscience Institute, Princeton University, United States

Corresponding author: Radulescu, Angela (angelar@princeton.edu)

Current Opinion in Neurobiology 2019, **55**:160–166

This review comes from a themed issue on **Machine learning, big data, and neuroscience**

Edited by **Jonathan Pillow** and **Maneesh Sahani**

For a complete overview see the Issue and the Editorial

Available online 30th April 2019

https://doi.org/10.1016/j.conb.2019.03.011

Introduction

Mental illness is often accompanied by models of the external world that deviate from the norm, and by changes in the dynamics of information processing relative to these models. In the past two decades, the machine-learning framework of reinforcement learning has emerged as an exceptionally good theory of human and animal learning, shedding light on computational processes implemented by networks of neurons in the service of decision making. Reinforcement-learning algorithms rely on models of the external world — a representation of the environment as a set of states that transition to one another given different actions, and that can generate rewards or punishments. Reinforcement-learning theory, therefore, naturally links aberrant models of the world to measurable changes in behavior, and can help quantify their neurocomputational substrates. This approach first emerged with neural network models of context maintenance deficits in schizophrenia [1–4]. Current frontiers include reinforcement learning theories of mood disorders [5–7,8••], obsessive-compulsive disorder [9•], and anxiety disorders [10,11].

Reinforcement learning models of mood disorders

Mood can be broadly defined as a valenced affective state that persists over longer periods of time [12]. Starting from the premise that mood can be understood as a feature of the brain's valuation system, recent reinforcement learning models have formalized mood as a running average of reward prediction errors [8••]. These theories focus on the valence of mood, treating mood as a scalar value that ranges from positive to negative, and seek to predict its behavioral and neural correlates. For example, expected value and prediction errors, but not total earnings, correlate with happiness ratings collected while participants make choices between different gambles [6]. Moreover, the extent to which reward prediction errors can be decoded from neurophysiological markers predicts self-reported mood fluctuations over the course of a week [13].

Eldar *et al.* further suggest that mood fluctuations can arise due to bidirectional coupling between mood and reinforcement learning: positive prediction errors increase mood, and positive mood increases the subjective value of rewards (and vice versa for negative mood), leading to a positive feedback loop that leaves the agent vulnerable to mood instability (Box 1). This provides a unified neurocomputational framework for both unipolar and bipolar mood disturbances [14•]. For instance, unipolar depression could arise by assuming that the impact of mood on reinforcement learning is asymmetric: negative mood may bias rewards more strongly than positive mood. Faster mood updates in the negative direction would also predict persistent unipolar depression. In contrast, a symmetric bidirectional interaction, can result in persistent fluctuations of mood, as in cyclothymia or bipolar disorder [8••].

The framework described above attributes mood disturbances to an altered interaction between mood and learning from prediction errors. Another key component of reinforcement learning is the state — a subset of features in the environment relevant to one's current goal. It is often implicitly assumed that states directly correspond to percepts, but in a complex world, states may be only 'partially observable' [15,16]. As a result, humans must construct an appropriate state representation by inferring which features are currently relevant [17], and supplementing sensory features with relevant components recalled from memory of past events [18]. A compact state representation that includes all relevant information and is free from extraneous information can render otherwise intractable decisions trivial [19]. For example, when deciding whether to hire a candidate or not, information about how they performed at their previous job is a better predictor of future success than features such as the color of their shirt or their packing habits. Correctly defining state is critical to both model-free and model-based reinforcement learning algorithms, as both rely on states to learn appropriate policies [20].

Important for reinforcement learning theories of mood disorders is that state representation can influence prediction errors in two ways: 1) by changing which states we form reward expectations over — for example, we might be more disappointed not to get a promotion if we knew the criteria and that we fulfilled them all, than if we had more doubt about our eligibility; and 2) by changing which states we attribute outcomes to — over time we may realize that the stated criteria for promotion are often waived, and management instead promotes those who repeatedly request promotions, even if not completely eligible. This would not only affect our policy in interacting with our superiors, but also the prediction errors we experience in our job. We next review evidence for the idea that given the behavioral demands of learning and decision-making in multidimensional environments, different manifestations of mental illness may be distinguished by changes in state representation.

Biased state representation along the bipolar spectrum

Bipolar disorder is a psychiatric condition in which patients experience mood disturbances in the form of manic and depressive episodes, interleaved with periods of neutral mood (euthymia). Bipolar disorder is distinguished from unipolar depression by the presence of one or more manic (bipolar-I) or hypomanic (bipolar-II) episodes. Core characteristics of mania include euphoria, elevated psychomotor agitation, and goal-directed activity, which may also be accompanied by more risk-taking, increased self-esteem and irritability [21]. Bipolar disorder is particularly prevalent among vulnerable populations, and carries a substantial risk of suicide. Patients with bipolar disorder are more likely to seek treatment when in the depressed rather than manic state, and as a consequence, bipolar disorder can be easily misdiagnosed as unipolar depression [22]. Cognitive theories of mood disorders have largely focused on depression [5,23,24•]. Here, we examine analogous theories of bipolar disorder, and suggest ways in which they may be cast in the language of state representation for reinforcement learning.

One prominent finding is that bipolar disorder is associated with a greater willingness to expend effort when pursuing reward [25], a pattern that stands in direct contrast to

Box 1 State representation effects on mood dynamics

Reinforcement learning models provide a useful computational language for formalizing the interaction between affect and behavior. One fundamental problem in reinforcement learning is determining a suitable *state representation* for the problem at hand: in addition to learning which actions to take, an agent must also learn which features are relevant for predicting reward, that is, which features ought to be learned about.

In general, when multiple features are present (e.g. features A and B), the total reward expectation (value V) can be computed as a weighted sum of expectations associated with component features:

$$V[AB] = \Phi_A V[A] + \Phi_B V[B],$$

where Φ are (learned) attention weights to each of the features. The predicted reward contingent on each feature can then be updated based on the difference between the perceived reward ($R_{perceived}$) and expected value, that is, the reward prediction error $PE = R_{perceived} - V[AB]$, scaled by a learning rate η, and again weighted by attention:

$$V_{new}[A] = V_{old}[A] + \eta \cdot PE \cdot \Phi_A$$
$$V_{new}[B] = V_{old}[B] + \eta \cdot PE \cdot \Phi_B.$$

In recent reinforcement-learning models of mood, reward perception was further biased by a mood-dependent term [7,8••]: $R_{perceived} = f \cdot m + R_{actual}$, where $m \in [-1,1]$ is mood and f is an individual trait parameter that quantifies mood reactivity. In this model, for positive values of f, positive mood biases the perception of outcomes upwards, and negative mood biases downwards, and vice versa for negative values of f. Mood itself is thought to reflect a running average of recent reward prediction errors:

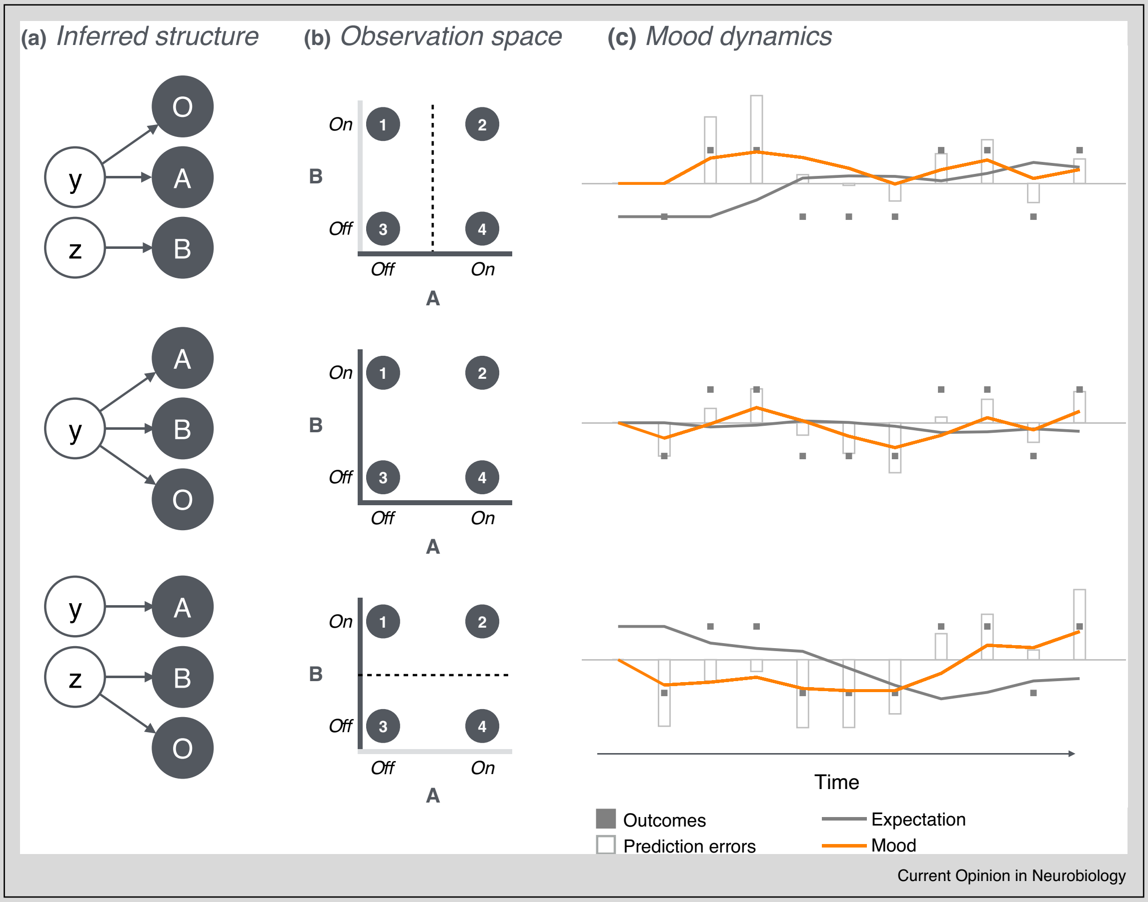
$h_{new} = h_{old} + \eta' \cdot (PE - h_{old})$ history of recent reward prediction errors

$m = \tanh(h)$ mood is the history of prediction errors mapped to $[-1, 1]$

Eldar *et al.* [7,8••] showed that this coupling between mood and reward can lead to positive-feedback dynamics that trigger cyclic mood fluctuations even in relatively stable environments, and that the degree of coupling is predictive of self-reported hypomanic tendencies.

How are the relative weights Φ determined? One way to think of these is as an index of the allocation of attention to each feature, which should follow from the structure of the environment [19]. For example, if the agent has inferred that a single latent variable y causes both A and the reward outcome (Figure 1a, top), she should ignore B ($\Phi_A = 1$, $\Phi_B = 0$), since only the presence (or absence) of A is relevant for predicting the outcome (Figure 1b, top). In this case, behavior should be determined based on distinctions along feature A alone. In contrast, if a common cause has generated A, B, and the outcome (Figure 1a, middle), then A and B should both be included in the state representation (Figure 1b, middle), potentially receiving equal attention ($\Phi_A = 0.5$, $\Phi_B = 0.5$) when computing and updating expectations, as each of them informs inference about the possible presence of the latent variable (and thus predictions about the outcome itself). When features have different initial reward expectations (here, $V_0[A] = -1$ and $V_0[B] = 1$), different state representations can lead to different mood dynamics (Figure 1c, compare to [8••], Figure 2a).

Figure 1



(a) Possible inferred latent structures lead to **(b)** different distinctions in observation space. Shaded circles correspond to observations, whereas empty circles are the underlying unobservable states or latent causes. If the latent cause is 'on' it tends to emit its linked observations. When a single perceptual feature A is related to the outcome via a common latent cause y (a, top), the agent can ignore B when computing reward expectations and deciding what action to take based on the resulting prediction. That is, observations 1 and 3 should be treated as equivalent and incorporated into a single state 'A off', and similarly for observations 2 and 4 that correspond to the state 'A on' (b, top). If both A and B are related to the outcome via a common cause, attention should be allocated to both A and B, as the presence of each provides information about the likelihood of *y* being active (b, middle). **(c)** Mood dynamics simulated over 10 timesteps, with both A and B present and the outcome (+1) or (−1) with equal probability. Initial values: $V_0[A] = -1$ and $V_0[B] = 1$, update rate parameters: $\eta = 0.25$, $\eta' = 0.5$ for value and mood respectively, mood reactivity: $f = 1.5$. A bias to attend to A throughout leads to low initial expectations, and thus elevated mood on average due to the positive prediction errors (c, top). Conversely, because of high initial expectations, attending to B leads high initial expectations and a dip in mood (c, bottom). Attention to both leads to milder mood fluctuations (c, middle).
Figure adapted with permission from Ref. [19].

findings in depression [26]. For instance, individuals diagnosed with bipolar disorder were faster to complete a card sorting task than controls when given the opportunity to earn reward [27]. This putative increase in approach motivation is associated with abnormally high striatal activity when anticipating reward, even during the euthymic state [28–31]. These findings are consistent with the notion that bipolar disorder may be accompanied by a generalized belief about reward being more abundant in the environment, and could explain why patients with bipolar disorder experience larger prediction errors and, perhaps as a result, more severe mood fluctuations.

But how may this belief arise in the first place? One possibility is that bipolar disorder involves changes in state representation that lead to optimistic expectations about future reward. People prone to mania may interpret the environment as being in a state where good things happen, and exhibit greater behavioral approach activity as a consequence. Indeed, manic symptoms tend to be associated with self-reported optimistic self-schemas. Individuals diagnosed with bipolar disorder who are more likely to perceive themselves as possessing positive self-dispositional traits (e.g. creative, successful) show increased rates of relapse of hypomania or mania [32].

Moreover, people diagnosed with bipolar disorder are more likely to become overconfident after experiencing success, even when performing tasks in which success occurs randomly, independent of their actions [33,34••,35]. This pattern of beliefs and actions could result from altered state representation. For example, when considering a simple coin toss, one may attribute outcomes to a single latent cause — a fair coin that randomly falls on heads or tails. Alternatively, one may construct the state space by inferring separate latent causes, one which favors heads (the 'lucky' state) and another which favors tails (the 'unlucky' state). Biasing state interpretation toward states that predict success (e.g. assuming that one is more likely to be in the 'lucky' state), possibly in a manner congruent with global optimistic self-schemas, would manifest in overconfidence.

Optimistic self-schemas in mania tend to be not only positively skewed, but also more extreme. In a self-report study assessing attributional styles, individuals with a history of bipolar disorder who made more extreme optimistic attributions (e.g. believing a good thing definitely happened due to one's actions, and will influence all situations in one's life) experienced significantly more lifetime episodes of mood elevation. It is thus possible that mania may be characterized by an overgeneralized positive outlook on the causes of events [36,37]. This view accords with a separate body of literature that has documented the effect of positive mood on the broadening of attention in healthy populations [12,38,39]. Taken together, these findings point to altered state representation in bipolar disorder, whereby individuals prone to mania form inflated, and more generalized prior reward expectations, perhaps based on optimistic prior beliefs about oneself.

Mood can, in turn, influence state representation by focusing learning toward certain stimuli in a mood-congruent manner. Such biases are well documented in unipolar depression, which is characterized by persistent selective attention to negative information from both memory and the external world [23,40–42]. Similar biases exist in bipolar disorder. Mansell and Lam found that people diagnosed with bipolar-I in a euthymic state are less likely to heed advice after a positive mood induction [43]. In this study, participants had to guess which of two choices led to a reward, and advice was operationalized as a computerized face providing the explicit reward probability. Participants in the bipolar group followed advice on significantly fewer trials than controls and people diagnosed with unipolar depression, suggesting that they were attending to different sources of information following the positive mood induction.

Such subtle differences between unipolar depression and bipolar disorder also manifest in the affective content of ruminative patterns. Patients diagnosed with bipolar disorder engage in rumination about negative emotion during depressive episodes, but they also focus on positive emotions during remission [33,44]. Moreover, the extent to which people ruminate over positive emotions is correlated with hypomania in healthy populations [45,46]. Individuals with bipolar disorder thus display mood-congruent state representation, focusing on negative information when in a bad mood, and positive information when in a good mood [47].

The link between elevated mood and optimistic overgeneralization is in line with a recent account that casts the interaction between mood and reward as a way to generalize learning across correlated sources of reward [8••]. Broadening attention in response to mood is an appropriate response if reward in the environment is actually correlated across multiple states. But if too many features become candidates for predicting reward and get incorporated into the state in a mood-congruent manner, expectations can become overly optimistic. When this overgeneralization is inconsistent with the actual statistics of reward in the environment, repeated thwarting of approach-related goals may lead to anger and irritability, both common symptoms of mania [21,35,48].

An important open question is whether altered state representation in mood disorders can be quantified as a tendency to attend to particular stimulus features when learning in multidimensional environments [49], and whether this tendency manifests as a trait or varies with mood. For example, it is possible that in bipolar disorder optimistic self-schemas may manifest as a mood-congruent positive bias in computing expectations: people prone to mania may be more likely to direct their attention to reward-predicting features after experiencing a sudden mood increase. Major depression patients, on the other hand, may display a tendency to focus on low-reward features regardless of mood. Extending current reinforcement learning models of mood dynamics to account for state representation phenomena would allow for a precise characterization of such patterns, and could provide a useful transdiagnostic marker for predicting illness course in bipolar disorder and unipolar major depression [9•,50].

State construction in schizophrenia

While in mood disorders patients tend to interpret the world in a relatively coherent (if extreme and affect-congruent) way, in schizophrenia the internal representation an become a severely distorted version of the external environment. Schizophrenia symptoms include hallucinations (perceptual aberrations), delusions (false beliefs), and disorganized thinking and speech (DSM-V). Anecdotally, patients diagnosed with schizophrenia describe a shattered world — 'like a photograph that is torn in bits and put together again' [51]. Measurable disturbances in perceptual organization which manifest even outside of major psychotic episodes (e.g. in the prodromal phase) are potential neurocognitive markers of schizophrenia, and have been hypothesized to emerge from a reduced ability to impose top–down structure on environmental cues [52,53]. Recent theories elaborating on this hypothesis posit an overreliance on external information when making inferences [54••], at the expense of prior beliefs that can help curtail runaway interpretations [55,56•]. Powers *et al.* [54••] report a particularly compelling increased tendency among people who hear voices to overweight sensory priors in a hallucination induction paradigm, and a dissociation between non-psychotic and psychotic participants in recognizing higher order statistics in the environment [54••]. Taken together, these findings suggest the possibility that psychotic disorders are characterized by altered state representation.

In particular, one important constraint on constructing states is the assumption of parsimony, that is, the tendency to ascribe events to as few states as possible [57,58]. All else equal, we should strive to explain current observations using the latent causes we already know exist, rather than assume new ones [59]. For instance, hearing a loud noise outside on a rainy evening should be ascribed to thunder due to the raging storm, rather than assuming that a bomb went off in the neighborhood, or aliens have landed an aircraft nearby. In models of state construction, such a bias for simplicity is manifest through a parameter that defines how likely it is, a priori, that a completely new state will cause the next observation [59]. This parameter can change with time and experience with the world.

A wealth of evidence from a simple learning paradigm called 'latent inhibition' suggests that patients diagnosed with schizophrenia (or even individuals with high schizotypy) infer new states more readily when experiencing unexpected situations (for instance, when the environment changes). In this paradigm, a neutral stimulus (e.g. a tone) is first exposed repeatedly ('preexposure' phase). In a later stage, the stimulus is paired with a motivationally relevant outcome (e.g. a mild shock or a reward). Control participants show slower learning of the stimulus–outcome relationship, compared to a group that did not undergo preexposure, suggesting that they are reluctant to assume a new state (e.g. one which consistently generates both the tone and the shock) absent ample evidence. Patients diagnosed with schizophrenia do not show this inhibition of learning [60], suggesting they readily ascribe the new information to a new state [59]. Animal models of drug-induced or lesion-induced schizophrenia mimic this behavior, and also show that typical and atypical antipsychotic drugs can reverse these effects [60]. Together with state interpretation that relies too heavily on observed events, this promiscuity in assuming new states can lead to a cycle of aberrant learning that would generate overly complex and somewhat baroque interpretations of reality.

Conclusion

Inferring useful representations of the environment is a key prerequisite for learning how to make decisions in a complex world. We have reviewed evidence for how state representation in reinforcement learning may be altered along the bipolar spectrum and in schizophrenia. One promise of this approach lies in carving a transdiagnostic space of neurocomputational constructs that can be charted by collecting large amounts of human behavioral data [9•,50]. By measuring individual differences in this space, the hope is not only to distinguish each person in need of treatment along dimensions useful for individualized diagnosis and management, but also to think across different levels of analysis and model systems, and better understand the etiology of mental illness at both the cognitive and neurobiological level [61].

Conflict of interest statement

Nothing declared.

Acknowledgements

We are grateful to Daniel Bennett and Eugenia Radulescu for comments on earlier drafts of the manuscript, and to Eran Eldar and Sam Zorowitz for helpful discussions. This work was supported by grant W911NF-14-1-0101 from the Army Research Office.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

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

2. Braver TS *et al.*: **Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function**. *Biol Psychiatry* 1999, **46**:312-328.

3. Waltz JA *et al.*: **Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction**. *Biol Psychiatry* 2006, **62**:756-764 http://dx.doi.org/10.1016/j.biopsych.2006.09.042.

4. Frank MJ: **Schizophrenia: a computational reinforcement learning perspective**. *Schizophr Bull* 2008, **34**:1008-1011.

5. Huys QJM *et al.*: **Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis**. *Biol Psychiatry* 2013, **3**:12.

6. Rutledge RB *et al.*: **A computational and neural model of momentary subjective well-being**. *Proc Natl Acad Sci U S A* 2014, **111**:12252-12257.

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

8. •• Eldar E *et al.*: **Mood as representation of momentum**. *Trends Cogn Sci* 2015, **20**:15-24.

Eldar *et al.* propose that mood can be understood in neurocomputational terms as a normative solution to inferring the overall momentum of recent outcomes. Mood is suggested to approximate momentum by tracking a running average of recent prediction errors and biasing the perception of outcomes. Through the coupling between mood and learning, agents can adapt to correlation structure in environmental rewards and generalize learning across multiple states. Dysfunctions in this system can lead to mood disturbances akin to those observed in mood disorders.

9. Gillan CM *et al.*: **Characterizing a psychiatric symptom dimension related to deficits in goal- directed control**. *eLife* 2016, **5**.
• This paper presents an approach for precisely quantifying compulsivity as a transdiagnostic computational construct using large-scale online data collection and computational modeling of behavior.

10. Mkrtchian A *et al.*: **Modeling avoidance in mood and anxiety disorders using reinforcement learning**. *Biol Psychiatry* 2017, **82**:532-539.

11. Huang H *et al.*: **Computational dysfunctions in anxiety: failure to differentiate signal from noise**. *Biol Psychiatry* 2017, **82**:440-446.

12. Fredrickson BL: **The broaden-and-build theory of positive emotions**. *Philos Trans R Soc Lond B Biol Sci* 2004, **359**:1367-1377.

13. Eldar E *et al.*: **Decodability of reward learning signals predicts report decodability of reward learning signals predicts mood fluctuations**. *Curr Biol* 2018, **28**:1-7.

14. Bennett D, Niv Y: **Detailed comparative study of current reinforcement learning models of major depressive disorder and bipolar disorder prefaced by a brief history of developments in computational psychiatry.**. *Opening Burton's Clock: Psychiatric Insights from Computational Cognitive Models*. 2018 http://dx.doi.org/10.31234/osf.io/y2vzu.
•

15. Kaelbling LP *et al.*: **Planning and acting in partially observable stochastic domains**. *Artif Intell* 1998, **101**:99-134.

16. Langdon AJ *et al.*: **Model-based predictions for dopamine**. *Curr Opin Neurobiol* 2018, **49**:1-7.

17. Wilson RC, Niv Y: **Inferring relevance in a changing world**. *Front Hum Neurosci* 2012, **5**:1-14.

18. Schuck NW *et al.*: **Human orbitofrontal cortex represents a cognitive map of state space**. *Neuron* 2016, **91**:1402-1412.

19. Radulescu A *et al.*: **Holistic reinforcement learning: the role of structure and attention**. *Trends Cogn Sci* 2019, **23**:278-292 http://dx.doi.org/10.1016/j.tics.2019.01.010.

20. Moran R *et al.*: **Retrospective model-based inference guides model-free credit assignment**. *Nat Commun* 2019, **10**:750.

21. Goodwin FK *et al.*: *Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression*. Oxford University Press; 2007.

22. Singh T, Rajput M: **Misdiagnosis of bipolar disorder**. *Psychiatry* 2006, **3**:57-63.

23. Gotlib IH, Joormann J: **Cognition and depression: current status and future directions**. *Annu Rev Clin Psychol* 2010, **6**:285-312.

24. Huys QJM *et al.*: **Depression: a decision-theoretic analysis**. *Annu Rev Neurosci* 2015, **38**:1-23.
• A thorough review of depression through the lens of Bayesian Decision Theory, mapping experimental findings to different computational hypotheses about how depressive behavior may arise from failures in the brain's valuation system.

25. Johnson SL *et al.*: **The double-edged sword of goal engagement: consequences of goal pursuit in bipolar disorder**. *Clin Psychol Psychother* 2012, **19**:352-362.

26. Whitton AE *et al.*: **Reward processing dysfucntion in major depression bipolar disorder and schizophrenia**. *Curr Opin Psychiatry* 2015, **28**:7-12.

27. Hayden EP *et al.*: **A multimethod investigation of the behavioral activation system in bipolar disorder**. *J Abnorm Psychol* 2008, **117**:164-170.

28. Lawrence NS *et al.*: **Subcortical and ventral prefrontal cortical neural responses to facial expressions distinguish patients with bipolar disorder and major depression**. *Biol Psychiatry* 2004, **55**:578-587.

29. Hassel S *et al.*: **Elevated striatal and decreased dorsolateral prefrontal cortical activity in response to emotional stimuli in euthymic bipolar disorder: no associations with psychotropic medication load**. *Bipolar Disord* 2008, **10**:916-927.

30. Nusslock R *et al.*: **Waiting to win: elevated striatal and orbitofrontal cortical activity during reward anticipation in euthymic bipolar disorder adults**. *Bipolar Disord* 2012, **14**:249-260.

31. Caseras X *et al.*: **Ventral striatum activity in response to reward: differences between bipolar I and II disorders**. *Am J Psychiatry* 2013, **170**:533-541.

32. Lam D *et al.*: **Sense of hyper-positive self and response to cognitive therapy in bipolar disorder**. *Psychol Med* 2005, **35**:69-77.

33. Eisner LR *et al.*: **Cognitive responses to failure and success relate uniquely to bipolar depression versus mania**. *J Abnorm Psychol* 2008, **117**:154-163.

34. Meyer TD *et al.*: **Vulnerability factors for bipolar disorders as predictors of attributions in ability-based and chance-based tests**. *J Individ Diff* 2010, **31**:29-37.
•• Meyer *et al.* show that people who are prone to mania make more self-serving attributions about positive feedback in a chance-based dice-throwing task, a pattern of beliefs that could result from altered state representation when learning from outcomes.

35. Johnson SL *et al.*: **The behavioral activation system and mania**. *Annu Rev Clin Psychol* 2012, **8**:243-267.

36. Stange JP *et al.*: **Positive overgeneralization and Behavioral Approach System (BAS) sensitivity interact to predict prospective increases in hypomanic symptoms: a behavioral high-risk design**. *Behav Res Ther* 2012, **50**:231-239.

37. Stange JP *et al.*: **Extreme attributions predict transition from depression to mania or hypomania in bipolar disorder**. *J Psychiatr Res* 2013, **47**:1329-1336.

38. Fredrickson BL, Branigan C: **Positive emotions broaden the scope of attention and thought-action repertoires thought-action repertoires**. *Cogn Emot* 2005, **19**:313-332.

39. Wadlinger HA, Isaacowitz DM: **Positive mood broadens visual attention to positive stimuli**. *Motiv Emot* 2006, **30**:87-99.

40. Blaney PH: **Affect and memory: a review**. *Psychol Bull* 1986, **99**:229-246 American Psychological Association.

41. Gotlib IH *et al.*: **Attentional biases for negative interpersonal stimuli in clinical depression**. *J Abnorm Psychol* 2004, **113**:127-135.

42. Mathews A, MacLeod C: **Cognitive vulnerability to emotional disorders**. *Ann Rev Clin Psychol* 2005, **1**:167-195.

43. Mansell W, Lam D: **"I Won't Do What You Tell Me!": elevated mood and the assessment of advice-taking in euthymic bipolar I disorder**. *Behav Res Ther* 2006, **44**:1787-1801.

44. Gruber J *et al.*: **Hooked on a feeling: rumination about positive and negative emotion in inter-episode bipolar disorder**. *J Abnorm Psychol* 2011, **120**:956-961.

45. Feldman GC *et al.*: **Responses to positive affect: a self-report measure of rumination and dampening**. *Cognit Ther Res* 2008, **32**:507-525.

46. Raes F *et al.*: **A psychometric evaluation of the Dutch version of the responses to positive affect questionnaire**. *Psychol Belg* 2010, **49**:293.

47. Alloy LB *et al.*: **Cognitive vulnerability to bipolar spectrum disorders**. *Cognitive Vulnerability to Emotional Disorders*. 2005:93-124.

48. Carver CS, Harmon-Jones E: **Anger is an approach-related affect: evidence and implications**. *Psychol Bull* 2009, **135**:183-204.

49. Leong YC *et al.*: **Dynamic interaction between reinforcement learning and attention in multidimensional environments**. *Neuron* 2017, **93**:451-463.

50. Gillan CM, Whelan R: **What big data can do for treatment in psychiatry**. *Curr Opin Behav Sci* 2017, **18**:34-42.

51. Chapman J: **The early symptoms of schizophrenia**. *Br J Psychiatry* 1966, **112**:225-251.

52. Silverstein SM, Ulhaas PJ: **Gestalt psychology: the forgotten paradigm in abnormal psychology**. *Am J Psychol* 2004, **117**:259-277.

53. Silverstein SM, Keane BP: **Perceptual organization impairment in schizophrenia and associated brain mechanisms: review of research from 2005 to 2010**. *Schizophr Bull* 2011, **37**:690-699.

54. Powers AR *et al.*: **Pavlovian conditioning-induced**
•• **hallucinations result from overweighting of perceptual priors**. *Science* 2017, **600**:596-600.
Powers *et al.* report that pairing a stimulus in one modality (vision) with a stimulus in another (sound) can lead to task-induced hallucinations in healthy individuals: after many trials, people eventually report perceiving a nonexistent stimulus contingent on the presence of the previously paired stimulus. Individuals who are more prone to hearing voices were more susceptible to these induced auditory hallucinations, possibly due to overweighting sensory information when representing state.

55. Denève S, Jardri R: **Circular inference: mistaken belief, misplaced trust**. *Curr Opin Behav Sci* 2016, **11**:40-48.

56. Jardri R *et al.*: **Experimental evidence for circular inference in**
• **schizophrenia**. *Nat Commun* 2017, **8**.
By exposing patients with schizophrenia and healthy controls to a task in which participants had to combine prior evidence with the likelihood of the current observation, the authors show that schizophrenia is associated with increased confidence in perceptual decisions, and that confidence scales with the amount of sensory data and ignores prior information. This study provides preliminary evidence that an overreliance on sensory evidence may mediate psychotic symptoms in schizophrenia.

57. Gershman SJ, Niv Y: **Perceptual estimation obeys Occam's razor**. *Front Psychol* 2013, **4**:623.

58. Gershman SJ *et al.*: **Discovering latent causes in reinforcement learning**. *Curr Opin Behav Sci* 2015, **5**:43-50.

59. Gershman SJ *et al.*: **Context, learning, and extinction**. *Psychol Rev* 2010, **117**:197-209.

60. Weiner I: **The "two-headed" latent inhibition model of schizophrenia: modeling positive and negative symptoms and their treatment**. *Psychopharmacology* 2003, **169**:257-297.

61. Fried EI, Nesse RM: **Depression sum-scores don't add up: why analyzing specific depression symptoms is essential**. *BMC Med* 2015, **13**:72.