37 Opening Burton's Clock: Psychiatric Insights from Computational Cognitive Models

DANIEL BENNETT AND YAEL NIV

ABSTRACT Computational psychiatry is a nascent field that seeks to use computational tools from neuroscience and cognitive science to understand psychiatric illness. In this chapter we make the case for computational cognitive models as a bridge between the cognitive and affective deficits experienced by those with a psychiatric illness and the neurocomputational dysfunctions that underlie these deficits. We first review the history of computational modeling in psychiatry and conclude that a key moment of maturation in this field occurred with the transition from qualitative comparison between computational models and human behavior to formal quantitative model fitting and model comparison. We then summarize current research at one of the most exciting frontiers of computational psychiatry: reinforcement-learning models of mood disorders. We review state-of-the-art applications of such models to major depression and bipolar disorder and outline important open questions to be addressed by the coming wave of research in computational psychiatry.

> The brain must needs primarily be misaffected, as the seat of reason … for our body is like a clock, if one wheel be amiss, all the rest are disordered; the whole fabric suffers.
> —Robert Burton, *The Anatomy of Melancholy*

For a watch repairer, the first task in fixing a faulty watch is diagnosis: What is the dysfunctional mechanism that is responsible for the fault? If the watch is losing time, is it because the mainspring is insufficiently wound, or could dirt be causing the gears to stick? If the watch has stopped, could this be the result of a loose balance wheel, or does the battery simply need changing?

In his analogy between human mental illness and the faulty mechanics of a clock, Robert Burton captured the essence of one of the most durable problems of contemporary biological psychiatry. In a clock a given functional disturbance, such as running fast or running slow, may be the result of any number of mechanical faults, and it is typically impossible to determine which mechanism is primarily amiss by observing the timekeeping dysfunction alone. Moreover, this inverse problem grows in difficulty with the complexity of the mechanism inside the watch: a fault is easier to diagnose when the underlying mechanism is simpler (e.g., a vibrating quartz crystal in a modern analog watch) than when it is complex (e.g., the many gears and springs of a 17th-century watch). Analogously, it has long been understood that psychiatric symptoms such as thought disorder and mania are aberrant behaviors produced by dysfunctions within an exceedingly complex dynamical system, the human brain (Hoffman, 1987; Joseph, Frith, & Waddington, 1979). It is no surprise, then, that identifying the specific neural-processing deficits that cause a given psychiatric symptom is difficult.

In this chapter we argue that computational psychiatry should approach this problem using computational cognitive models, with a focus on testing specific behavioral predictions made by different candidate neurocomputational dysfunctions. Just as the ticking sounds of a clock can be decomposed with spectral analyses to diagnose a mechanical fault (He, Su, & Du, 2008), computational cognitive models can be used to infer the latent neurocomputational deficits that underlie psychiatric conditions as diverse as depression and psychosis. However, just as in the clock analogy, the utility of these inferences critically depends upon two factors: first, an accurate mechanistic model of how the system operates and second, a sensitive behavioral assay of its operations. To this end, computational psychiatry should seek to integrate normative and process models from computational neuroscience and biological psychiatry with behavioral tests from cognitive psychology, computer science, and economics. By applying computational cognitive models to sensitive measures of human behavior, we may make substantial progress in identifying the dysfunctions of neural computation that give rise to psychiatric illness.

This chapter first reviews the history of the computational-modeling paradigm in psychiatry through the cognitive revolution of the 1960s and 1970s and the rise of parallel distributed processing and

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

reinforcement-learning models in the 1980s and 1990s. We then summarize the current state of the art of computational psychiatry in the study of mood disorders such as major depression and bipolar disorder using reinforcement-learning models.

The History of Computational Psychiatry

> Psychopathology has been rather a disappointment to the instinctive materialism of the doctors, who have taken the view that every disorder must be accompanied by actual lesions of some specific tissue involved… . This distinction between functional and organic disorders is illuminated by the consideration of the computing machine.
>
> —Norbert Wiener, *Cybernetics*

The idea that psychiatric illness might result from dysfunctions of neural or mental computation was proposed within 10 years of the invention of the modern digital computer. Writing in 1948 as part of a broader argument that the central nervous system ought to be treated as a self-regulating circuit, Norbert Wiener suggested a novel perspective on the 19th-century psychiatric distinction between organic and functional disorders (Fürstner, 1881, as cited by Beer, 1996). This dichotomy contrasts organic disorders caused by a purely biological pathology (such as a brain tumor or neurodegeneration) with functional disorders that cannot be diagnosed solely by the inspection of brain tissue. Wiener proposed that functional disorders—among which he included schizophrenia and bipolar disorder—could be best understood by analogy with the operations of a computer. This was, he proposed, because deficits in these disorders arose not from aberrations in the physical structure of the brain but from dysfunctions in the way the physical structure processed information (Wiener, 1948).

This information-processing paradigm was immensely influential in early cognitive psychology but gained traction much more slowly in psychiatry. Early research using computational models in psychiatry was rudimentary and consisted of little more than qualitative comparisons between simple computational models and aspects of contemporary psychiatric theory. For instance, Callaway (1970) pursued the analogy of a malfunctioning computer in an attempt to understand conceptual disorganization and the loosening of associations in schizophrenia. Drawing upon contemporary advances in cognitive science, Callaway posited that cognitive structures in schizophrenia could be represented as simple computational architectures called TOTE (test-operate-test-exit) units (Miller, Galanter, & Pribram, 1960). Deficits in schizophrenia were posited to result from interference in the test operations of these units by excessive neural noise. While the TOTE architecture has not proved durable, Callaway's notion that deficits in schizophrenia result from excessive levels of noise in neural computation has remained influential to the present day (e.g., Silverstein, Wibral, & Phillips, 2017; Winterer & Weinberger, 2004).

Separately, Colby (1964) used a computational dictionary seeded with quotations from human psychiatric patients to generate synthetic dialogues resembling those of a therapist with a psychiatric patient (e.g., "Father preferred sister. I avoid father." Colby, 1964, p. 221). Colby proposed that distorted beliefs in psychosis arose as a result of conflict between mutually exclusive impulses. Colby, Hilf, Weber, and Kraemer (1972) presented practicing psychotherapists with teletype printouts of a number of putative therapist/patient dialogues—half real and half generated by algorithm—and assessed the therapists' ability to distinguish real patients from simulated ones. It was found that therapists could not identify the real patients at an above-chance level and in some cases offered detailed psychoanalytic interpretations of the unconscious processes underlying algorithmically generated dialogues. The algorithm that generated the text engaged in dialogue by performing a rudimentary form of natural language processing with the intention of classifying its interlocutor's statements as either malevolent, benevolent, or neither. Depending on the values of the variables used to perform this classification, the algorithm then selected an internal response (e.g., anger or fear) and a corresponding utterance (e.g., verbal hostility in the case of high levels of anger). This algorithm can therefore be thought of as an early cognitive model of psychosis (albeit one that does not invoke unconscious processing, contrary to then-dominant theoretical ideas).

Other early work applying computational and mathematical methods to psychiatric illness did not adapt the computer metaphor directly. For instance, Rashevsky (1964) posited a rudimentary biophysical neural-processing system to explain the positive symptoms of schizophrenia in terms of the excessive reinforcement of endogenously generated responses. Houghton (1969) sought to specify a formal mathematical framework for understanding psychoanalysis by positing a negative feedback relationship between an "id module" and an "ego module," resulting in distortions of a topological space. Such theories have little empirical relevance for contemporary research; instead, they primarily reinforce the importance of grounding models of psychiatric illness in biologically principled models of neural computation.

The first computational models that are of more than historical interest to current research in computational

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

psychiatry were made possible by advances in computational models of neural information processing. For instance, a computational theory of the distribution of attention among stimuli based on recurrent lateral inhibition between noisy processing channels (Walley & Weiden, 1973) gave rise directly to a computational model of attentional deficits in schizophrenia (Joseph, Frith, & Waddington, 1979). This model proposed that an excess of dopaminergic activity led to increased overall levels of mutual inhibition between sensory inputs in schizophrenia and thereby to a dysfunction in the system's ability to produce winner-take-all network dynamics.

The advent of more advanced neural network architectures in the 1980s stimulated the development of more sophisticated computational psychiatric models. For instance, Hopfield (1982) described a fully interconnected neural network that produced emergent properties resembling human recognition memory, categorization, and generalization. In turn, Ralph Hoffman showed how dysfunctions of computation within Hopfield nets led to aberrant dynamics resembling schizophrenia and mania (Hoffman, 1987) and linked the putative computational deficit in schizophrenia to aberrant patterns of cortical pruning in frontal cortex (Hoffman & Dobscha, 1989). At the same time, the immense influence of parallel distributed-processing connectionist architectures in cognitive science (Rumelhart & McClelland, 1987) led naturally to the adaptation of multilayer neural networks for psychiatric research (e.g., Ruppin, 1995; Spitzer, 1995; Stein & Ludik, 1998).

Of particular note, Cohen and Servan-Schreiber (1992) used a multilayer neural network to model a failure to maintain mental context in schizophrenia. This work demonstrated a *quantitative* correspondence between the behavior of trained neural network models and the behavior of patients with schizophrenia on three tasks: a Stroop task, a continuous performance task, and a lexical disambiguation task. The computational mechanism by which these deficits were produced in the model was a reduction of the gain of units in the network representing task context, and this computational dysfunction was linked by the authors to decreased dopaminergic activity in the prefrontal cortex in schizophrenia. This work marks a point of transition between qualitative and quantitative comparisons of models and behavior in computational psychiatry. As such, it stands in contrast to prior research that had proceeded after the fashion of Callaway (1970) by suggesting *qualitative* parallels between patterns of information processing in psychiatric illness and patterns of information processing in real or hypothetical computational architectures.

Arguably, this development—the quantitative fitting of computational models to behavior produced by individuals with a psychiatric illness—is responsible for much of the subsequent achievement, and much of the future promise, of computational methods in psychiatric research. The ability of computational models to make quantitative predictions about human behavior means that different psychiatric theories can be compared by instantiating each as a different model and determining which model provides the most accurate and parsimonious account of behavior. Once identified, a model serves at least two purposes: First, it provides a quantitative device for the measurement of cognitive-psychiatric symptoms that may aid in diagnosis and treatment selection in psychiatry in much the same way that a blood glucose test aids in diagnosing and treating diabetes. Second, a good correspondence between the predictions of a model and observed behaviors may offer a window into the functional causes of aberrant experiences in psychiatric illness, since it suggests mechanisms by which these symptoms may be produced.

As computational approaches to psychiatry have expanded in recent years, the behavioral model-fitting and model-comparison paradigm has grown to encompass computational models from disciplines including economic game theory (King-Casas et al., 2008), hierarchical probabilistic inference (Friston, Stephan, Montague, & Dolan, 2014), and Bayesian decision theory (Huys, Daw, & Dayan, 2015). In the remainder of this chapter, we review these developments with a specific focus on the state-of-the-art computational modeling of two mood disorders: major depression and bipolar disorder. In particular, we explore the extent to which dysfunctions in these conditions can be understood through the lens of reinforcement learning (see, e.g., Maia & Frank, 2011).

Reinforcement Learning Models of Mood Disorders

Below, we summarize the insights that reinforcement-learning models provide into the neurocomputational substrates of depression and bipolar disorder. Our intention is not to claim that mood disorders are disorders of learning narrowly defined. Instead, we argue that the mathematical formalisms of reinforcement learning provide a language that can describe how representations of the reinforcement value of the environment go astray in mood disorders.

Briefly, reinforcement learning describes a set of computational principles by which an agent in an uncertain or complex environment can act to maximize future expected reward (Dayan & Niv, 2008; Sutton & Barto, 1998). The framework relies on several relatively

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

simple psychological primitives: Representations of different states of the environment, of the actions that can be taken by the agent in each state, and of the rewards that are received following each action. Reinforcement-learning algorithms then describe operations by which an agent can update its representations of the values of different actions as it interacts with the environment.

The foundational computational variable in reinforcement learning is the prediction error δ, calculated as the difference between the actual reward received after taking some action and the amount of reward an agent had expected to result from that action:

$$\delta = R_t - Q_t\,(s_t,\, a_t) \tag{37.1}$$

Here, R_t denotes the reward (or, if negative, punishment) received on trial t, and $Q_t\,(s_t,\, a_t)$ denotes the expected value on trial t of taking action a_t in state s_t. δ takes a positive value when the received reward exceeds the expected reward amount (a positive reward prediction error) and a negative value when the reward received is less than expected. Given this prediction error, one can then update expectations for trial $t+1$ according to a simple Rescorla-Wagner learning rule (Rescorla & Wagner, 1972):

$$Q_{t+1}\,(s_t,\, a_t) = Q_t(s_t,\, a_t) + \eta \cdot \delta \tag{37.2}$$

where η is a learning rate parameter controlling the speed with which action values are acquired. Equation 37.2 ensures that the expected value of actions will be incremented following positive reward prediction errors and decremented following negative reward prediction errors. Neurally, the prediction error signal δ (and, more precisely, its temporal difference cousin that accounts for the timing of prediction error signals within a trial; Schultz, Dayan, & Montague, 1997) is thought to be instantiated in the brain by the phasic release of dopamine in the basal ganglia.

From this foundation we can derive increasingly complex and sophisticated reinforcement-learning algorithms. For instance, the simple update rule described above is typically referred to as model-free reinforcement learning since it learns solely about the value of taking particular actions in particular states and not about the structure of the environment itself. This contrasts with model-based reinforcement learning, in which agents learn an internal model of the environment (possibly using prediction error signals) and use this model to plan actions through mental simulations of alternative options and their predicted outcomes (see, e.g., Doll, Simon, & Daw, 2012).

The domain of reinforcement learning is an agent's cognitive and behavioral responses to the affective feedback (i.e., rewards and punishments) that it receives from the environment. This domain is also a primary area of cognitive dysfunction in mood disorders, including major depression and bipolar disorder (Admon & Pizzagalli, 2015; Eshel & Roiser, 2010; Whitton, Treadway, & Pizzagalli, 2015). As such, reinforcement-learning models are well suited to the study of neurocomputational dysfunction in mood disorders. For instance, individuals with depression show a number of cognitive biases consistent with a reduced learned value of the environment and the preferential processing of negative information, such as pessimistic expectations regarding the value of future events (Showers & Ruben, 1990), an increased tendency to retrieve negatively valenced items from memory (Blaney, 1986), and decreased sensitivity to rewarding feedback (Henriques & Davidson, 2000). Similarly, a recent theory has suggested that oscillatory mood dynamics characteristic of bipolar disorder might be produced by an interaction between mood and the valuation of outcomes (Eldar & Niv, 2015; Eldar, Rutledge, Dolan, & Niv, 2016). As we will show, each of these phenomena can be described well in terms of dysfunctions of computation within a reinforcement-learning model.

Depression

Phenomenology and theories of depression The two most common diagnostic taxonomies of psychiatric illness, the Diagnostic and Statistical Manual of Mental Disorders (*DSM-5*) and the International Statistical Classification of Diseases (*ICD-10*), concur on two primary symptoms of major depression: persistent low mood or sadness and an inability to take pleasure in everyday events (anhedonia). The two taxonomies also concur on other secondary symptoms of depression, including fatigue or lack of energy (anergia), poor concentration, disturbances of sleep and appetite, thoughts of suicide or self-harm, feelings of guilt or worthlessness, and psychomotor disturbances (either agitation or motor slowing).

Cognitive theories of depression have posited a number of distinct information- processing biases that might underlie these symptoms (Gotlib & Joormann, 2010; Ingram, 1984). For instance, Beck (1967) proposed that preexisting representations (*schemas*) of oneself, other people, and the external world bias the processing of emotional information in a schema-congruent way. One example of a depressive schema, for instance, is a core belief that one is unlovable; this belief would lead to the interpretation of neutral or ambiguous social cues as consistent with the fact that one is unlovable, thereby reinforcing the schema. Other cognitive theories have emphasized the operation of different

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

cognitive processes, but most agree that the biased processing of emotional information plays a crucial role in the onset and maintenance of depression. For instance, Bower (1981) and Ingram (1984) emphasized the role of disturbed semantic networks in depression, leading to the increased activation of negatively valenced nodes in an associative network. By contrast, Lewinsohn (1974) adopted a behaviorist perspective and emphasized the role of a lack of response-contingent reinforcement in depression, whereas Rehm (1977) emphasized the role of self-control in the selective processing of negative outcomes, and Seligman (1975) highlighted the role of learned helplessness (that is, the distorted belief that one's experiences of positive and negative events are not under one's own control).

Cognitive theories of depression have been highly influential, both in empirical research on cognition in depression and in the development of applied cognitive therapies for depression. However, these theories are persistently criticized because they merely redescribe known phenomena and do not offer any novel insights (Blaney, 1977; Ingram, 1984). The computational approach to psychiatry that we argue for in this chapter provides a tool to address this shortcoming. This is because the requirement that theories of psychiatric illness be embedded in a computational model means that quantitative behavioral predictions of different theories can be generated directly via model simulation. Empirical work can then test the extent to which these predictions are borne out by human behavior. Additionally, by mapping information-processing biases in depression onto putative neural computations—especially within the framework of reinforcement learning—computational models can flesh out cognitive theories of depression with reference to our understanding of how these computations are implemented in the human brain.

Computational modeling of depression The basic reinforcement-learning framework detailed in equations 37.1 and 37.2 can be extended to capture the cognitive phenomena of depression in a number of ways. One possibility proposed by Huys, Pizzagalli, Bogdan, and Dayan (2013) is that anhedonia represents a diminished hedonic response to rewarding outcomes in depression, which affects prediction errors as below:

$$\delta = \rho \cdot R_t - Q_t(s_t, a_t) \tag{37.3}$$

where $0 \leq \rho \leq 1$ is a reward sensitivity parameter that describes the degree to which primary hedonic responses to rewarding outcomes are diminished in individuals with depression. The pattern of behavior produced by this model matches the phenomenological experience of anhedonia in the sense that since the effective reward value of outcomes is diminished, individuals with lower values of ρ will experience outcomes as subjectively less rewarding. Because reinforcement learning from prediction errors means they will also learn that the reward value of actions and options in the environment is lower, such individuals will form pessimistic expectations about future outcomes.

To provide evidence for this model, Huys et al. (2013) fit a version of the computational model described by equation 37.3 to the behavior of individuals with varying levels of anhedonia as they performed a simple learning task designed to measure reward sensitivity (Pizzagalli, Jahn, & O'Shea, 2005). Huys et al. (2013) found that across both healthy individuals and those with major depression, self-reported anhedonia was positively correlated with participants' estimated reward sensitivity ρ but not their estimated learning-rate parameter η.

However, further evidence complicates this view and suggests that anhedonia should not be simply viewed as a deficiency in hedonic responses to rewarding outcomes (Huys et al., 2015). If it were true that primary hedonic responses to rewards were diminished in depression, it would be expected that individuals with depression would report less enjoyment of pleasant primary rewards, such as sweet liquids. However, this is not the case: those with depression do not differ from healthy controls in the self-reported pleasantness of sucrose solutions (Amsterdam, Settle, Doty, Abelman, & Winokur, 1987). In addition, a recent study found no differences between those with depression and healthy controls in the strength of the relationship between reward prediction error magnitude and self-reported mood during a gambling task (Rutledge et al., 2017). This leads to the question: What computational mechanisms other than reduced hedonic response to rewards might explain an apparent reduction in reward sensitivity in depression?

A re-examination of cognitive theories of depression suggests asymmetric responses to positive and negative outcomes as one candidate. For instance, the self-control theory of Rehm (1977) proposes that depression is associated with selective attention to negative outcomes, as well as a tendency to make stronger inferences about the self from negative feedback than positive feedback. Similarly, the reinforcement theory of Lewinsohn (1974) posits that a reduction in the degree to which actions are reinforced by positive feedback is central to depression. From the perspective of reinforcement learning, one way of capturing this proposed information-processing bias is as an asymmetry in learning rates for positive versus negative reward

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

prediction errors (Gershman, 2015; Mihatsch & Neuneier, 2002; Niv, Edlund, Dayan, & O'Doherty, 2012):

$$Q_{t+1}(s_t,\ a_t) = \begin{cases} Q_t(s_t,\ a_t) + \eta^+ \cdot \delta, & \delta > 0 \\ Q_t(s_t,\ a_t) + \eta^- \cdot \delta, & \delta < 0 \end{cases} \quad (37.4)$$

In equation 37.4, η^+ is the learning rate for positive reward prediction errors, and η^- is the learning rate for negative reward prediction errors. When $\eta^- > \eta^+$, value updates are affected more strongly by negative reward prediction errors, consistent with the proposed negative information-processing bias in major depression. This bias produces an underestimation of the value of uncertain rewards that is qualitatively similar to that produced by a reduction of the reward sensitivity parameter ρ in equation 37.3. However, deterministic rewards are learned correctly by this model (Niv et al., 2012).

Importantly, underestimations of reward value could be produced in equation 37.4 by hypersensitivity to negative reward prediction errors (increased η^-), by hyposensitivity to positive reward prediction errors (decreased η_+), or both. Empirical evidence from behavioral studies of depression is divided on this question. While there is consistent evidence that individuals with depression display diminished learning from positive feedback (Henriques & Davidson, 2000; Henriques, Glowacki, & Davidson, 1994; Korn, Sharot, Walter, Heekeren, & Dolan, 2014; Robinson, Cools, Carlisi, Sahakian, & Drevets, 2012; Vrieze et al., 2013), evidence for increased sensitivity to negative feedback is more equivocal. Some studies have shown that those with depression respond more to worse than expected outcomes than healthy controls, (Garrett et al., 2014; Nelson & Craighead, 1977) but others have found no difference (Henriques & Davidson, 2000; Henriques, Glowacki, & Davidson, 1994; Robinson et al., 2012; Santesso et al., 2008). This suggests, on balance, that aberrant reward processing in depression is more likely to result from hyposensitivity to positive reward prediction errors than from hypersensitivity to negative reward prediction errors. Further study of this question is required, however, and an important open question is whether different symptom profiles of depression are associated with different patterns of learning from positive and negative reward prediction errors. For instance, it is known that anxiety, a disorder highly comorbid with major depression (Sartorius, Üstün, Lecrubier, & Wittchen, 1996), is associated with hypersensitivity to punishment and increased attention to potentially threatening events (Bishop, 2007). This suggests the interesting possibility that low-level computational mechanisms of depression might differ between major depression with and without comorbid anxiety.

As a further prediction, asymmetric learning rates as per equation 37.4, but not changes in reward sensitivity as per equation 37.3, induce preferences with respect to the *risk* of outcomes (in the economic sense of risk, referring to outcome variance; Mihatsch & Neuneier, 2002). Learning rate asymmetry in depression would therefore also predict that individuals with depression should display increased risk aversion. This is because high-risk choice options are those associated with larger deviations, on average, between individual instances of reward and long-term reward averages, meaning larger absolute reward prediction errors. As a result, high-risk choice options will be more devalued when $\eta^- > \eta^+$ than low-risk choice options, resulting in risk aversion. This prediction is consistent with behavioral data showing increased risk aversion in individuals with depression performing the Iowa Gambling Task (Smoski et al., 2008), as well as greater self-reported risk aversion (Leahy, Tirch, & Melwani, 2012; Wiersma et al., 2011).

Separately, recent theories in computational psychiatry have also proposed a role for the dysfunction of model-based reinforcement learning in depression. As introduced above, model-based reinforcement learning applies to scenarios in which an agent's decisions are dependent upon a learned internal model of the environment (a *model* of the environment, hence *model-based* reinforcement learning). This is distinguished from model-free reinforcement learning, in which agents learn solely about the values of individual actions (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). Two candidate model-based mechanisms for depression proposed by Huys et al. (2015) are biased attention toward negative possibilities in internal estimates of a current state and a failure to "prune" negative states from contemplation in planning future sequences of action.

The first of these, a bias in the internal representation of a state, reflects the fact that states of the world (s in the equations above) are not necessarily observable features; instead, a "state" represents an agent's inferences about the structure of rewards in the world at a given point in time and about the way that structure may change if different actions are taken (Schuck, Cai, Wilson, & Niv, 2016). For instance, while waiting at a bus stop, one can only estimate whether the state of the world is "the bus is shortly arriving" or "the bus already passed and I missed it." If the inferences used to construct this state are biased in a pessimistic way—such as because negative potential outcomes are weighted more strongly than positive outcomes—then an agent may believe itself to be in a worse state than is truly the case. Such a process might underlie the pessimistic representations of future outcomes in depression and might also provide an explanation for experiences of

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

anergia, since low response vigor and reduced energy expenditure are rational strategies for an agent to adopt in states where few rewarding outcomes can result from action.

The second model-based mechanism is a failure to "prune" negative states from future planned actions in depression. In planned decision-making, nondepressed individuals typically avoid excessive focus upon the future possible states associated with large negative outcomes (Huys et al., 2012). This is an adaptive strategy since it means that cognitive resources can be directed instead toward plans that have a high a priori chance of reaching future states associated with a high reward value. Less pruning of negative states would be associated with a relatively greater focus on negative-valued paths in future planning, potentially leading to the patterns of ruminative thought characteristic of depression (Whitmer & Gotlib, 2013).

Open questions for the computational modeling of depression The literature reviewed above suggests several important open questions to be addressed via the computational modeling of behavior in depression.

First, to what extent can anhedonia in depression be characterized by asymmetric learning from positive and negative reward prediction errors, rather than reduced consummatory pleasure in reward receipt? Second, what combination of model-based and model-free reinforcement learning best describes the cognitive deficits observed in depression? On the one hand, depression may be associated with a low-level asymmetry in (model-free) learning. On the other hand, depression may be better characterized by model-based deficits in the construction of the present state and planning for future states. Or, depression may involve both deficits. Importantly, these questions can be answered using computational models and tasks specifically tailored to measure the parameters of these models in each individual.

Finally, how might the computational deficits underlying depression be expressed in different contexts? As Beck (1967) observed, inferences in depression are far more likely to be negatively biased when their object is one's own worth than when their object is an abstract statistical quantity. In the language of reinforcement learning, it is almost certainly not the case that learning rates for positive and negative prediction errors will be expressed equivalently in all domains. Instead, one possibility is that individual differences in the allocation of attention to positive and negative outcomes in different settings might provide a principled explanation for apparent differences in reinforcement sensitivity in depression. For instance, it is possible that attention to outcomes—and therefore learning rates—may fluctuate commensurate with the outcomes' congruency with prior beliefs regarding oneself. Designing sensitive measures of the context-dependence of reinforcement learning dysfunction in depression is therefore a crucial task for future research.

Bipolar Disorder

Phenomenology and subtypes of bipolar disorder In contrast to major depression, which is characterized solely by episodes of depression, bipolar disorder is characterized by episodes of both depression and mania. Under common definitions in the *DSM-5* and *ICD-10*, *mania* refers to a state in which mood is elevated (euphoria), and there is increased energy and goal-directed activity. Mania, and its less severe counterpart hypomania, are also typically characterized by increased risk-taking behavior, a decreased subjective need for sleep, and increased self-esteem, potentially leading to delusions of grandiosity (Goodwin & Jamison, 2007).

Typologies of bipolar disorder distinguish between two subtypes, bipolar I and bipolar II, which differ in the relative frequency and intensity of manic and depressed episodes. Bipolar I disorder is characterized by at least one episode of mania and often (but not necessarily) by other episodes of depression. By contrast, bipolar II disorder is typified by episodes of both major depression and hypomania (not meeting the full criteria for mania). Both forms of bipolar disorder are typified by a functional recovery between episodes of mania or depression to a mood in the normal range.

Whereas cognitive theories of depression have abounded since the 1960s, until recent years bipolar disorder was largely viewed through a psychopharmacological lens (Goodwin & Jamison, 2007), with a relative paucity of cognitive theorizing (but see, e.g., Alloy et al., 2008). One finding in this literature, however, is of mood-congruent information-processing biases in bipolar disorder. That is, individuals with bipolar disorder may display negative information-processing biases when in a low mood, as in depression, but positive information-processing biases when in a good mood (for reviews, see Alloy, Reilly-Harrington, Fresco, & Flannery-Schroeder, 2005; Whitton, Treadway, & Pizzagalli, 2015). This mood congruence is a critical feature of bipolar disorder that computational models must seek to account for; it also represents a significant point of contrast with cognitive theories of depression, which rather emphasize trait-level information-processing biases as a cognitive mechanism for the disorder.

Computational modeling of bipolar disorder A recent model has posited a set of computational mechanisms that

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

may partly explain mood-congruent information-processing biases in bipolar disorder. Using a reinforcement-learning framework, Eldar and Niv (2015) proposed that mood oscillations and information-processing biases may be governed by a dynamic interaction between mood and outcome valuation. Specifically, their model proposed that the reward value of outcomes R_t is biased by a mood-dependent factor f^{m_t} in the calculation of prediction errors:

$$\delta = f^{m_t} \cdot R_t - Q_t(s_t, a_t) \tag{37.5}$$

Here, $-1 \leq m_t \leq 1$ represents mood at trial t, with negative values of m_t denoting negatively valenced moods and positive values of m_t denoting positively valenced moods. f is a parameter governing the strength of the interaction between mood and outcome valuation such that values of f greater than 1 indicate mood-congruent changes in outcome valuation (i.e., the overestimation of outcome value in good moods and the underestimation of outcome value in bad moods).

The model also proposes that mood changes over time according to a weighted average of recent reward prediction errors that is transformed to lie between −1 and 1 by a sigmoidal function:

$$h_{t+1} = h_{t+1} + \eta_\eta \cdot (\delta - h_t) \tag{37.6}$$

$$m_t = \tanh(\delta - h_t) \tag{37.7}$$

where η_h is a learning-rate parameter for this reward prediction error history. Together, equations 37.5–37.7 specify a dynamic system in which reward prediction errors trigger the mood-congruent processing of subsequent rewards. This, in turn, leads to escalatory mood dynamics that may explain the emergence of mania and depression in bipolar disorder.

There is an important parallel between this model of bipolar disorder and the models of depression reviewed above. Specifically, the form of equation 37.5 closely resembles that of the reward-sensitivity model of depression in equation 37.3, as posited by Huys et al. (2013). The difference between the two models is that Huys et al. (2013) posit a trait-level parameter ρ to govern blunted reward sensitivity in depression, whereas Eldar and Niv (2015) propose a mood-dependent term f^{m_t}.

This comparison may be instructive. In reviewing the models of depression above, we observed that the reward-sensitivity model of depression posited by Huys et al. (2013) made predictions similar to a model in which depression affected not the hedonic value of rewards (through ρ) but rather the asymmetry between the effects of positive and negative reward prediction errors (through η^+ and η^-). A similar principle applies to models of bipolar disorder. This means that an alternative model to that of Eldar and Niv (2015) is one in which mood affects not the hedonic value of rewards but the relative strength of learning from positive versus negative reward prediction errors:

$$Q_{t+1}(s_t, a_t) = \begin{cases} Q_t(s_t, a_t) + f^{m_t} \cdot \eta^+ \cdot \delta, & \delta > 0 \\ Q_t(s_t, a_t) + f^{-m_t} \cdot \eta^- \cdot \delta, & \delta < 0 \end{cases} \tag{37.8}$$

where δ is defined according to equation 37.3, not equation 37.5. The cognitive interpretation of equation 37.8 is that positive moods lead to increases in learning rate from positive reward prediction errors and decreases in learning from negative reward prediction errors and vice versa for negative moods.

Here, too, the reward-sensitivity model of Eldar and Niv (2015) and the model specified by equation 37.8 make different predictions concerning attitudes toward risk in bipolar disorder. This is because equation 37.8, but not the model of Eldar and Niv (2015), predicts that positive moods should be associated with decreased risk aversion (increased risk seeking). This is consistent with a large body of evidence suggesting that mania and hypomania are associated with increased risk-taking behavior (e.g., Mason, O'Sullivan, Montaldi, Bentall, & El-Deredy, 2014; Thomas, Knowles, Tai, & Bentall, 2007), as well as with diagnostic guidelines specifying risk-taking as a symptom of bipolar disorder in the *DSM-5*. Testing this prediction via behavioral model fitting in bipolar disorder is therefore a key task for future research.

Conclusion

In the 17th century, Robert Burton compared psychiatric illness to a clock in which one faulty gear interfered with the operation of the whole machine. In adapting this metaphor, we realize that in every age the brain has been likened to the most sophisticated contemporary machine—including clocks, steam locomotives, and now digital computers—none of which the brain is likely all that similar to. Nevertheless, the present chapter has considered how, given such a clock, we might apply computational methods to determine which gear is at fault. We have reviewed the history of a computational approach to psychiatric illness, with a focus on the current state of the art for reinforcement-learning models of major depression and bipolar disorder. Cutting-edge future research in this field will involve two lines of work: research to identify the algorithmic principles that govern human mood and affect and research to characterize how these algorithms go awry in psychiatric illness. Our contention is that these questions are best addressed by adapting computational cognitive models to human behavioral data.

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

A strong version of our behavioral argument holds that it is only by making distinct predictions about human behavior that psychiatric theories can meaningfully differ from one another. After all, if two different psychiatric theories made entirely equivalent predictions about behavior (and therefore about all phenomenological aspects of a patient's experience that are accessible to empirical inquiry), it would be reasonable to conclude that these two theories were functionally isomorphic, even if they proposed seemingly dissimilar theoretical constructs to explain psychiatric dysfunction (Putnam, 1975). A less strong, more pragmatic version of this same argument is that by adopting the quantitative prediction of behavior as the ground truth of psychiatric theory, it is relatively straightforward to reject theories that may seem conceptually sound while making no sensible predictions regarding behavior (e.g., Houghton, 1969). A focus on the prediction of behavior evaluates theories according to their empirical content and not the sophistication of their mathematical superstructures.

If it is true that scientific revolutions occur not necessarily because of serendipitous discovery but because certain scientists come to ask better questions, then the promise of computational psychiatry lies in the nature of the questions that it can ask about psychiatric illness. We propose that as a source for such questions, computational cognitive models are a critically important tool. Such models can be used to identify the nature of the computations employed by the brain, the role of aberrant computations in the production of psychiatric illness, and the potential biological and cognitive remedies for computational dysfunction.

Acknowledgment

This work was supported by a CJ Martin Early Career Fellowship (#1165010) to DB from the NHMRC.

REFERENCES

Admon, R., & Pizzagalli, D. A. (2015). Dysfunctional reward processing in depression. *Current Opinion in Psychology, 4*, 114–118.

Alloy, L. B., Abramson, L. Y., Walshaw, P. D., Cogswell, A., Grandin, L. D., Hughes, M. E., et al. (2008). Behavioral approach system and behavioral inhibition system sensitivities and bipolar spectrum disorders: Prospective prediction of bipolar mood episodes. *Bipolar Disorders, 10*(2), 310–322.

Alloy, L. B., Reilly-Harrington, N. A., Fresco, D. M., & Flannery-Schroeder, E. (2005). Cognitive vulnerability to bipolar spectrum disorders. In Lauren B. Alloy & John H. Riskind (Eds.), *Cognitive vulnerability to emotional disorders* (pp. 93–124). Hillsdale, NJ: Erlbaum.

Amsterdam, J. D., Settle, R. G., Doty, R. L., Abelman, E., & Winokur, A. (1987). Taste and smell perception in depression. *Biological Psychiatry, 22*(12), 1481–1485.

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. Philadelphia: University of Pennsylvania Press.

Beer, M. D. (1996). The dichotomies: Psychosis/neurosis and functional/organic: A historical perspective. *History of Psychiatry, 7*(26), 231–255.

Bishop, S. J. (2007). Neurocognitive mechanisms of anxiety: An integrative account. *Trends in Cognitive Sciences, 11*(7), 307–316.

Blaney, P. H. (1977). Contemporary theories of depression: Critique and comparison. *Journal of Abnormal Psychology, 86*(3), 203.

Blaney, P. H. (1986). Affect and memory: A review. *Psychological Bulletin, 99*(2), 229.

Bower, G. H. (1981). Mood and memory. *American Psychologist, 36*(2), 129.

Burton, R. (1847). *The anatomy of melancholy*. New York: Wiley and Putnam. (Original work published 1621.)

Callaway, E. (1970). Schizophrenia and interference: An analogy with a malfunctioning computer. *Archives of General Psychiatry, 22*(3), 193–208.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review, 99*(1), 45–77.

Colby, K. M. (1964). Experimental treatment of neurotic computer programs. *Archives of General Psychiatry, 10*(3), 220–227.

Colby, K. M., Hilf, F. D., Weber, S., & Kraemer, H. C. (1972). Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence, 3*, 199–221.

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*(6), 1204–1215.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology, 18*(2), 185–196.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology, 22*(6), 1075–1081.

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

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences, 20*(1), 15–24.

Eshel, N., & Roiser, J. P. (2010). Reward and punishment processing in depression. *Biological Psychiatry, 68*(2), 118–124.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *Lancet Psychiatry, 1*(2), 148–158.

Fürstner, C. (1881). Über delirium acutum. *Archiv für Psychiatrie und Nervenkrankheiten, 11*, 517–531.

Garrett, N., Sharot, T., Faulkner, P., Korn, C. W., Roiser, J. P., & Dolan, R. J. (2014). Losing the rose tinted glasses: Neural substrates of unbiased belief updating in depression. *Frontiers in Human Neuroscience, 8*, 639.

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review, 22*(5), 1320–1327.

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

Goodwin, F. K., & Jamison, K. R. (2007). *Manic-depressive illness: Bipolar disorders and recurrent depression*. Oxford: Oxford University Press.

Gotlib, I. H., & Joormann, J. (2010). Cognition and depression: Current status and future directions. *Annual Review of Clinical Psychology, 6*, 285–312.

He, Q., Su, S., & Du, R. (2008). Separating mixed multicomponent signal with an application in mechanical watch movement. *Digital Signal Processing, 18*(6), 1013–1028.

Henriques, J. B., & Davidson, R. J. (2000). Decreased responsiveness to reward in depression. *Cognition & Emotion, 14*(5), 711–724.

Henriques, J. B., Glowacki, J. M., & Davidson, R. J. (1994). Reward fails to alter response bias in depression. *Journal of Abnormal Psychology, 103*(3), 460.

Hoffman, R. E. (1987). Computer simulations of neural information processing and the schizophrenia-mania dichotomy. *Archives of General Psychiatry, 44*(2), 178–188.

Hoffman, R. E., & Dobscha, S. K. (1989). Cortical pruning and the development of schizophrenia: A computer model. *Schizophrenia Bulletin, 15*(3), 477–490.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*(8), 2554–2558.

Houghton, G. (1969). A lie group topology for normal and abnormal human behavior. *Bulletin of Mathematical Biophysics, 31*(2), 275–293.

Huys, Q. J., Daw, N. D., & Dayan, P. (2015). Depression: A decision-theoretic analysis. *Annual Review of Neuroscience, 38*, 1–23.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology, 8*(3), e1002410.

Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: A behavioral meta-analysis. *Biology of Mood & Anxiety Disorders, 3*(1), 12.

Ingram, R. E. (1984). Toward an information-processing analysis of depression. *Cognitive Therapy and Research, 8*(5), 443–477.

Joseph, M. H., Frith, C. D., & Waddington, J. L. (1979). Dopaminergic mechanisms and cognitive deficit in schizophrenia. *Psychopharmacology, 63*(3), 273–280.

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*(5890), 806–810.

Korn, C., Sharot, T., Walter, H., Heekeren, H., & Dolan, R. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine, 44*(3), 579–592.

Leahy, R. L., Tirch, D. D., & Melwani, P. S. (2012). Processes underlying depression: Risk aversion, emotional schemas, and psychological flexibility. *International Journal of Cognitive Therapy, 5*(4), 362–379.

Lewinsohn, P. M. A. (1974). A behavioral approach to depression. In R. J. Friedman & M. M. Katz (Eds.), *The psychology of depression: Contemporary theory and research* (pp. 157–184). Washington, DC: V. H. Winston.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience, 14*(2), 154.

Mason, L., O'Sullivan, N., Montaldi, D., Bentall, R. P., & El-Deredy, W. (2014). Decision- making and trait impulsivity in bipolar disorder are associated with reduced prefrontal regulation of striatal reward valuation. *Brain, 137*(8), 2346–2355.

Mihatsch, O., & Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine Learning, 49*(2–3), 267–290.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Henry Holt.

Nelson, R. E., & Craighead, W. E. (1977). Selective recall of positive and negative feedback, self-control behaviors, and depression. *Journal of Abnormal Psychology, 86*(4), 379.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience, 32*(2), 551–562.

Pizzagalli, D. A., Jahn, A. L., & O'Shea, J. P. (2005). Toward an objective characterization of an anhedonic phenotype: A signal-detection approach. *Biological Psychiatry, 57*(4), 319–327.

Putnam, H. (1975). Philosophy and our mental life. In *Philosophical papers vol. 2: Mind, language, and reality (*pp. 291–303). Cambridge: Cambridge University Press.

Rashevsky, N. (1964). A neurobiophysical model of schizophrenias and of their possible treatment. *Bulletin of Mathematical Biophysics, 26*(2), 167–185.

Rehm, L. P. (1977). A self-control model of depression. *Behavior Therapy, 8*(5), 787–804.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (Vol. 2, pp. 64–99). New York: Appleton-Century-Crofts.

Robinson, O. J., Cools, R., Carlisi, C. O., Sahakian, B. J., & Drevets, W. C. (2012). Ventral striatum response during reward and punishment reversal learning in unmedicated major depressive disorder. *American Journal of Psychiatry, 169*(2), 152–159.

Rumelhart, D. E., & McClelland, J. L. (1987). *Parallel distributed processing* (Vol. 1). Cambridge, MA: MIT Press.

Ruppin, E. (1995). Neural modelling of psychiatric disorders. *Network: Computation in Neural Systems, 6*(4), 635–656.

Rutledge, R. B., Moutoussis, M., Smittenaar, P., Zeidman, P., Taylor, T., Hrynkiewicz, L., … Dolan, R. J. (2017). Association of neural and emotional impacts of reward prediction errors with major depression. *JAMA Psychiatry, 74*(8), 790–797.

Santesso, D. L., Steele, K. T., Bogdan, R., Holmes, A. J., Deveney, C. M., Meites, T. M., & Pizzagalli, D. A. (2008). Enhanced negative feedback responses in remitted depression. *Neuroreport, 19*(10), 1045.

Sartorius, N., Üstün, T. B., Lecrubier, Y., & Wittchen, H.-U. (1996). Depression comorbid with anxiety: Results from the WHO study on "psychological disorders in primary health care." *British Journal of Psychiatry, 168*(S30), 38–43.

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron, 91*(6), 1402–1412.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599.

Seligman, M. E. P. (1975). *Helplessness: On depression, development, and death*. New York: W. H. Freeman.

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

Showers, C., & Ruben, C. (1990). Distinguishing defensive pessimism from depression: Negative expectations and positive coping mechanisms. *Cognitive Therapy and Research, 14*(4), 385–399.

Silverstein, S. M., Wibral, M., & Phillips, W. A. (2017). Implications of information theory for computational modeling of schizophrenia. *Computational Psychiatry, 1*, 82–101.

Smoski, M. J., Lynch, T. R., Rosenthal, M. Z., Cheavens, J. S., Chapman, A. L., & Krishnan, R. R. (2008). Decision-making and risk aversion among depressive adults. *Journal of Behavior Therapy and Experimental Psychiatry, 39*(4), 567–576.

Spitzer, M. (1995). A neurocomputational approach to delusions. *Comprehensive Psychiatry, 36*(2), 83–105.

Stein, D. J., & Ludik, J. (1998). *Neural networks and psychopathology: Connectionist models in practice and research*. Cambridge: Cambridge University Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Thomas, J., Knowles, R., Tai, S., & Bentall, R. P. (2007). Response styles to depressed mood in bipolar affective disorder. *Journal of Affective Disorders, 100*(1), 249–252.

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*(7), 639–645.

Walley, R. E., & Weiden, T. D. (1973). Lateral inhibition and cognitive masking: A neuropsychological theory of attention. *Psychological Review, 80*(4), 284–302.

Whitmer, A. J., & Gotlib, I. H. (2013). An attentional scope model of rumination. *Psychological Bulletin, 139*(5), 1036.

Whitton, A. E., Treadway, M. T., & Pizzagalli, D. A. (2015). Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. *Current Opinion in Psychiatry, 28*(1), 7.

Wiener, N. (1948). Cybernetics. *Scientific American, 179*(5), 14–19.

Wiersma, J. E., van Oppen, P., Van Schaik, D., Van der Does, A., Beekman, A., & Penninx, B. (2011). Psychological characteristics of chronic depression: A longitudinal cohort study. *Journal of Clinical Psychiatry, 72*(3), 288–294.

Winterer, G., & Weinberger, D. R. (2004). Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends in Neurosciences, 27*(11), 683–690.

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026

Downloaded from http://direct.mit.edu/books/edited-volume/chapter-pdf/2053702/c031600_9780262356176.pdf by Princeton University Library user on 13 January 2026