Review

Special Issue: Cognition in Neuropsychiatric Disorders

Computational psychiatry

P. Read Montague[1,2], Raymond J. Dolan[2], Karl J. Friston[2] and Peter Dayan[3]

[1] Virginia Tech Carilion Research Institute and Department of Physics, Virginia Tech, 2 Riverside Circle, Roanoke, VA 24016, USA
[2] Wellcome Trust Centre for Neuroimaging, University College London, 12 Queen Square, London, WC1N 3BG, UK
[3] Gatsby Computational Neuroscience Unit, Alexandra House, 17 Queen Square, London, WC1N 3AR, UK

Computational ideas pervade many areas of science and have an integrative explanatory role in neuroscience and cognitive science. However, computational depictions of cognitive function have had surprisingly little impact on the way we assess mental illness because diseases of the mind have not been systematically conceptualized in computational terms. Here, we outline goals and nascent efforts in the new field of computational psychiatry, which seeks to characterize mental dysfunction in terms of aberrant computations over multiple scales. We highlight early efforts in this area that employ reinforcement learning and game theoretic frameworks to elucidate decision-making in health and disease. Looking forwards, we emphasize a need for theory development and large-scale computational phenotyping in human subjects.

The explanatory gap

The idea of biological psychiatry seems simple and compelling: the brain is the organ that generates, sustains and supports mental function, and modern psychiatry seeks the biological basis of mental illnesses. This approach has been a primary driver behind the development of generations of anti-psychotic, anti-depressant, and anti-anxiety drugs that enjoy widespread clinical use. Despite this progress, biological psychiatry and neuroscience face an enormous explanatory gap. This gap represents a lack of appropriate intermediate levels of description that bind ideas articulated at the molecular level to those expressed at the level of descriptive clinical entities, such as schizophrenia, depression and anxiety. In general, we lack a sufficient understanding of human cognition (and cognitive phenotypes) to provide a bridge between the molecular and the phenomenological. This is reflected in questions and concerns regarding the classification of psychiatric diseases themselves, notably, each time the Diagnostic and Statistical Manual of Mental Disorders (DSM) of the American Psychiatric Association is revised [1].

While multiple causes are likely to account for the current state of affairs, one contributor to this gap is the (almost) unreasonable effectiveness of psychotropic medication. These medications are of great benefit to a substantial number of patients; however, our understanding of why they work on mental function remains rudimentary. For example, receptors are understood as molecular motifs (encoded by genes) that shuttle information from one cellular site to another. Receptor ligands, whose blockade or activation relieves psychiatric symptoms, furnished a kind of conceptual leap that seemed to obviate the need to account for the numerous layers of representation intervening between receptor function and behavioral change. This, in turn, spawned explanations of mental phenomena in simplistic terms that invoked a direct mapping from receptor activation to complex changes in mental status. We are all participants in this state of affairs, since symptom relief in severe mental disease is sufficient from a clinical perspective, irrespective of whether there are models that connect underlying biological phenomena to the damaged mental function. A medication that relieves or removes symptoms in a large population of subjects is

Glossary

Cognitive phenotype: a phenotype is a measureable trait of an organism. Although easy to state in this manner, the idea of a phenotype can become subtle and contentious. Phenotypes include different morphology, biochemical cascades, neural connection patterns, behavioral patterns and so on. Phenotypic variation is a term used to refer to those variations in some trait on which natural selection could act. A cognitive phenotype is a pattern of cognitive functioning in some domain that could be used to classify styles of cognition. By analogy, variations in cognitive phenotypes would be subject to natural selection.

Computational phenotyping: a computational phenotype is a measurable behavioral or neural type defined in terms of some computational model. By analogy with other phenotypes, a computational phenotype should show variation across individuals and natural selection could act on this variation. Large-scale computational phenotyping in humans has not been carried out; therefore, the ultimate utility of this idea has not been rigorously tested.

Game theory: the study of mathematical models of interactions between rational agents.

Instrumental controller: instrumental conditioning is the process by which reward and punishment are used in a contingent fashion to increase or decrease the likelihood that some behavior will occur again in the future. An instrumental controller is one whose control over behavior can be conditioned in exactly the same fashion. It is an operational term used in the reinforcement learning approach to motivated behavior to refer to any controller whose influence over behavior shows the dependence on rewards and punishments typical of instrumental conditioning.

Neuromodulatory systems: systems of neurons that project to broad regions of target neural tissue to modulate subsequent neural responses in those regions. Neuromodulatory systems typically have cell bodies situated in the brainstem and basal forebrain and deliver neurotransmitters, such as serotonin, dopamine, acetylcholine and norepinephrine, to target regions. They are called modulatory because their impact is typically much longer-lasting than fast synaptic effects mediated by glutamate and they are much more widely distributed.

Pavlovian controller: an operational name for a behavioral controller that is Pavlovian in the normal psychological use of this term – that is, the controller mediates involuntary responses to situations or stimuli. Pavlovian control can be demonstrated behaviorally and modern work is focused on identifying the neural substrates that contribute to this function.

Serotonin: a neuromodulator common to many neurons in the raphe nuclei. Serotonin has a presumed role in clinical depression because of the efficacy of medications that selectively block its reuptake into neurons after its release from synaptic terminals (so-called SSRI's – selective serotonin reuptake inhibitors).

Corresponding author: Montague, P.R. (read@vt.edu).

 1364-6613/$ – see front matter doi:10.1016/j.tics.2011.11.018

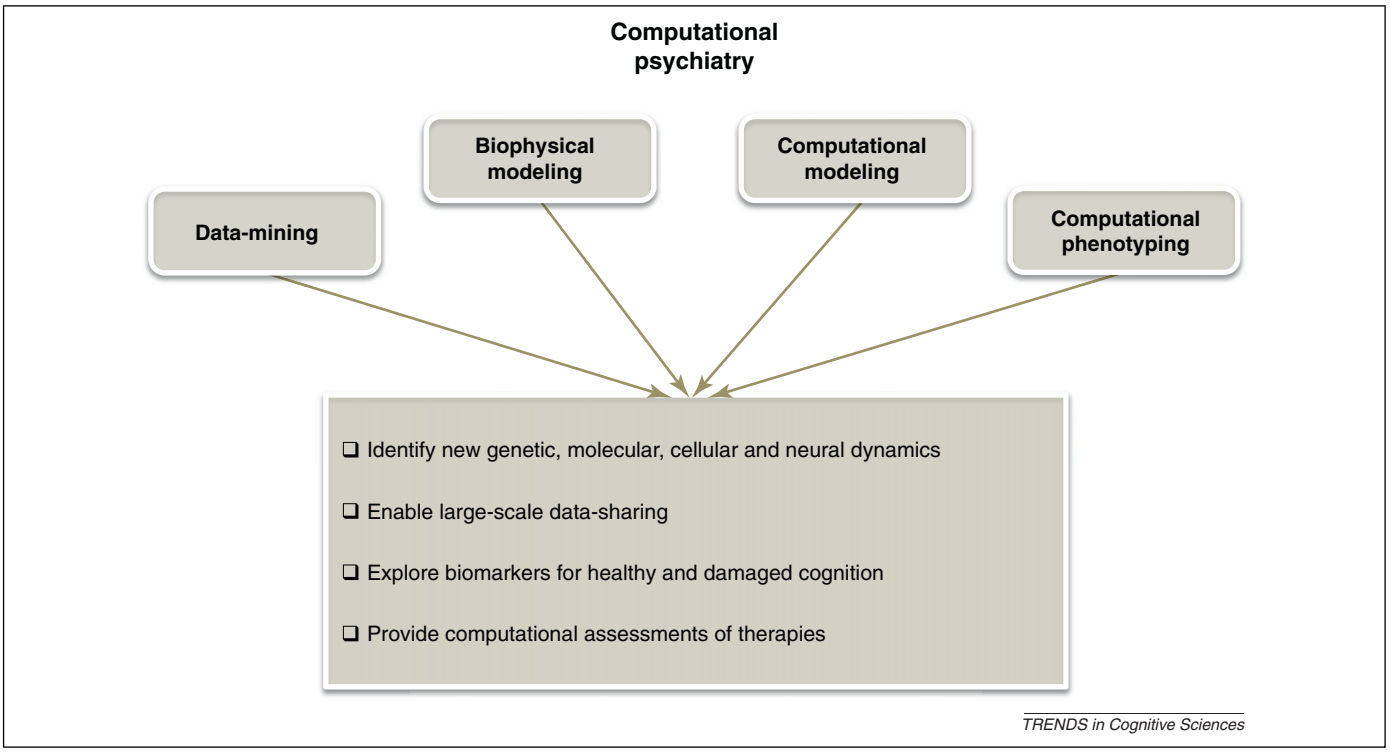

Figure 1. Components of Computational Psychiatry.

unquestionably of great utility, even if the explanation for why it works is lacking. However, significant gaps in the effectiveness of medications for different mental illness mean we should look to advances in modern neuroscience and cognitive science to deliver more.

We believe that advances in human neuroscience can bridge parts of the explanatory gap. One area where there has been substantial progress is in the field of decision-making. Aberrant decision-making is central to the majority of psychiatric conditions and this provides a unique opportunity for progress. It is the computational revolution in cognitive neuroscience that underpins this opportunity and argues strongly for the application of computational approaches to psychiatry. This is the basis of computational psychiatry [2–4] (Figure 1). In this article, we consider this emerging field and outline central challenges for the immediate future.

Contrasting mathematical and computational modeling

Mathematical modeling

To define computational modeling, we must first distinguish it from its close cousin, mathematical or biophysical modeling. Mathematical modeling provides a quantitative expression for natural phenomena. This may involve building multi-level (unifying) reductive accounts of natural phenomena. The reductions involve explanatory models at one level of description that are based on models at finer levels, and are ubiquitous in everything from treatments of action potentials [5] (see also [6] for a broader view) to the dynamical activity of populations of recurrently connected neurons [7]. Biophysical realism, however, is a harsh taskmaster, particularly in the face of incomplete or sparse data. For example, in humans, there seems to be little point in building a biophysically detailed model of the dendrite of single neurons if one can only measure synaptic responses averaged over millions of neurons and billions of synapses using functional magnetic resonance imaging (fMRI) or electroencephalography (EEG).

Biophysical modeling is important for elucidating key relationships in a hugely complex system [8] and thus predicting the possible effects of therapeutic interventions (see [9] for an example using dynamic causal modeling). For example, it is well known that critical mechanisms within neuromodulatory systems, such as dopamine, serotonin, norepinephrine and acetylcholine, are subject to intricate patterns of feedback and interactive control, with autoreceptors regulating the activity of the very neurons that release neuromodulators. Moreover, this feedback often includes the effects of one neuromodulator (e.g., serotonin) on the release and impact of others (e.g., dopamine) [10]. These neuromodulators are implicated in many psychiatric and neurological conditions. The fact that they play key roles in so many critical functions may explain the fact, if not the nature, of this exquisite regulation. It is the complexity of these interactions that invites biophysical modeling and simulation, for instance, to predict the effect of medication with known effects on receptors or uptake mechanisms. Moreover, the capacity to perform fast biophysical simulations is essential for evidence-based model comparison using empirical data [11] and the exploration of emergent behaviors (e.g., [12]). Simulation has become vital to vast areas of science and it will be central in computational psychiatry. Mathematical predictions based on real neural and biophysical data are important; however, they are not equivalent to a computational account of mental or neural function.

Computational modeling

Computational modeling seeks normative computational accounts of neural and cognitive function. Such accounts start from the premise that the brain solves computational problems and indeed has evolved to do so. One of the pioneers of computing theory, Alan Turing, conceived of mental function in exactly this fashion – the mind was cast as specific patterns of information processing supported by a particular kind of hardware (the brain) [13]. This notion implies key constraints on mental phenomena – in particular constraints on computational complexity that limit the power of any device, neural or mechanical, to solve a wide range of problems [14]. This idea is commonplace today but in the 1930s the idea of computation and its limits underwent a revolution [15–17] (see also [18]).

Currently, computational accounts of elements of mental and neural function exist, and in each case, typically some constraint is found that guides the discovery of the computational model. Some of the most important constraints come from optimality assumptions – the idea that the brain is organized to maximize or minimize quantities of external and internal importance (e.g., [6,19]). One set of optimality constraints emerges naturally from behaviors that support survival, such as foraging for food or responding appropriately to prospects of danger [20]. A wide range of ideas, proofs, methods and algorithms for executing such behaviors can be found in many fields, including engineering, economics, operations research, control theory, statistics, artificial intelligence and computer science. In fact, these fields provide a formal foundation for the interpretation of many cognitive and neural phenomena [6]. This foundation can span important levels of description, for instance, offering accounts of the representational semantics of the population activity of neurons [21] or of the firing of neuromodulatory neurons in the context of tasks involving predictions of reward [22–26]. This type of computational modeling can thus provide one explanatory framework for the reductive mathematical modeling discussed above.

Computational modeling in decision-making

The field of decision-making has been a particular target for computational modeling. Decision-making involves the accumulation of evidence associated with the utilities of possible options and then the choice of one of them, given the evidence. Decision-making problems in natural environments are extremely complex. One difficulty arises from the balance models must strike between built-in information acquired over the course of evolution about the nature of the decision-making environment (ultimate constraints) versus what can be learned over the course of moment-to-moment experience (proximate constraints). A second difficulty arises because of the inherent computational complexity of the problem: certain types of optimal decision-making appear intractable for any computational system. This fact motivates the search for approximations that underlie mechanisms actually used in animals. Reinforcement learning is one area where such approximations have been used to guide the discovery of neural and behavioral mechanisms. Box 1 provides a brief description of the modern view of neural reinforcement learning.

Many psychiatric conditions are associated not only with abnormal subjective states, such as moods, but also with aberrant decisions. Patients make choices: in depression, not to explore; in obsessive compulsive disorder, to repeat endlessly a behavior (such as hand-washing) that has no apparent basis in rational fact (such as having dirty hands); in addiction, to seek and take a drug, despite

Box 1. Reinforcement learning

Reinforcement learning (RL) is a field, partly spawned by mathematical psychology, that spans artificial intelligence, operations research, statistics and control theory (for a good introductory account of RL, see [89]). RL addresses how systems of any sort, be they artificial or natural, can learn to gain rewards and avoid punishments in what might be very complicated environments, involving states (such as locations in a maze) and transitions between states. The field of neural RL maps RL concepts and algorithms onto aspects of the neural substrate of affective decision-making [90,91]. One important feature of this framework is that the majority of its models can be derived from a normative model of how an agent 'should' behave under some explicit notion of what that agent is trying to optimize [89].

Conventional and neural RL include two very broad classes of method: model-based and model-free. Model-based RL involves building a statistical model of the environment (a form of cognitive map; see [92]) and then using it to (i) choose actions based on predicted outcomes and (ii) improve predictions by optimizing the model. Acquiring such models from experience can be enhanced by sophisticated prior expectations (a facet that we relate to the phenomena of learned helplessness). In other words, an agent significantly enhances the models it can build based on experience if it already starts with a good characterization of its environment. In turn, these models enable moment-to-moment prediction and planning. Except in very simple environments, prediction and planning consume enormous memory and computational resources – a fact has inspired much work on approximations and the search for biological work-arounds.

Model-free RL involves learning exactly the same predictions and preferences as model-based RL, but without building a model. Instead, model-free RL learns predictions about the environment by enforcing a strong consistency constraint: successive predictions about the same future outcomes should be the same. Actions are chosen based on the simple principle that actions which lead to better predicted outcomes are preferred. Model-free RL imposes much lower demands on computation and memory because it depends on past learning rather than present inference. However, this makes it less flexible to changes in the environment.

The conceptual differences between model-based and model-free RL suggest that correlates can be sought in real-world neural and behavioral data. There are ample results from animal and human experiments to suggest that both model-free and model-based RL systems exist in partially distinct regions of the brain [67,93–97] and that there is a rich panoply of competitive and cooperative interactions between them [67]. Model-free RL has a particularly close association with the activity of the dopamine neuromodulatory system, especially in the context of appetitive outcomes and predictions.

Finally, model-based and model-free RL are both instrumental in the sense that actions are chosen because of their consequences [94]. Animals are also endowed with extremely sophisticated Pavlovian controllers (see main text), where outcomes and predictions of those outcomes directly elicit a set of species-typical choices apparently not under voluntary control. One important example related to predictions of future negative outcomes is behavioral inhibition (learning not to do something), which may be related to serotonin [51].

explicitly acknowledging the damage that follows. Key to the initial form of computational psychiatry is the premise that, if the psychology and neurobiology of normative decision-making can be characterized and parameterized via a multi-level computational framework, it will be possible to understand the many ways in which decision-making can go wrong. However, we should first consider an important earlier tradition of modeling in psychiatry.

Early connectionist models of mental dysfunction

There is an old idea in brain science, namely, that complex functions emerge from networked interactions of relatively simple parts [27,28]. In the brain, the most conspicuous physical substrates for this idea are the networks of neurons connected by synapses. This perspective has been termed 'connectionism'. One modern expression of connectionism began with the work of Rumelhart, McClelland and the parallel distributed processing research group [29] (but now see [30]), which applied this approach to both brain and cognition in the early and mid-1980s, building upon the earlier pioneering work [27,28]. The basic concept underlying connectionism involves taking simple, neuron-like, units and connecting in them in ways that are either biologically plausible based on brain data or capable of performing important cognitive or behavioral functions. At approximately the same time and parallel to this work, three key publications emerged from physicists John Hopfield and David Tank, which showed how a connectionist-like network can have properties equivalent to those pertaining to the dynamics of a physical system [31–33]. Inspired by Hopfield's work and the seminal (and still classic) work of Stuart and Donald Geman on Gibbs sampling and Bayesian approaches to image analysis, Hinton and Sejnowski [34] showed that probabilistic activation in simple units could perform a sophisticated Bayesian style of inference. Collectively, this work addressed memory states, constraint satisfaction, pattern recognition and a host of other cognitive functions [29], thus suggesting that these models might aid in understanding mental disease.

Through the 1990s, connectionist models turned their sights on psychopathologies, such as schizophrenia [35–39]. These models primarily addressed issues related to cognitive control and neuromodulation [35–38], with a particular focus on neural systems that could support these functions [40–44]. These and other models offered plausible solutions for how networks of neurons could implement functions, such as cognitive control and memory, and offered new abstractions for how such functions go awry in specific pathologies. This work leans heavily on the neurally-plausible aspect of connectionist models, a feature that now finds more biological support, as neuroscience has produced enormous amounts of new data that can be fit into such frameworks [42–44].

Recent efforts toward computational characterization of mental dysfunction

In this section, we review recent efforts to develop and test computational models of mental dysfunction and to extract behavioral phenotypes relevant for building computationally-principled models of mental disease. The examples discussed are intended to provide insights into healthy mental function but in a fashion designed to inform the diagnosis and treatment of mental disease. Along with the pioneering earlier studies [35–40], there have been recent treatments and reports of work along these lines on schizophrenia [3,45,46], addiction [47], Parkinson's disease, Tourette's syndrome, and attention-deficit hyperactivity disorder [3]. Here, we concentrate on two areas that have not been recently reviewed in this context, namely depression and autism.

The efforts discussed here are now collectively blossoming into programmatic efforts in computational psychiatry (for example, the joint initiative of the Max Planck Society and University College London: Computational Psychiatry and Aging Research). It is our opinion that such efforts must reach further and strive to extract normative computational accounts of healthy and pathological cognition useful for building predictive models of individuals. Consequently, we emphasize for computational psychiatry the goal of extracting computational principles around which human cognition and its supporting biological apparatus is organized. Achieving this goal will require new types of phenotyping approaches, in which computational parameters are estimated (neurally and behaviorally) from human subjects and used to inform the models. This type of large-scale computational phenotyping of human behavior does not yet exist.

Reinforcement learning (RL) models of mood disorders and anxiety

Box 1 notes three different, albeit interacting, control systems within the context of RL: model-based, model-free, and Pavlovian. Model-based and model-free systems link the choice of actions directly to affective consequences. The Pavlovian system determines involuntary actions on the basis of predictions of outcomes, whether or not deployed actions are actually appropriate for gaining or avoiding those outcomes. Pavlovian control appears completely automated in this description. However, it is known that other brain systems can interact with Pavlovian control, hence, it is at this level that such control can be sensitive to ongoing valuations in other parts of the brain.

These types of controllers and their interactions have been the subject of computational modeling in the context of mood disorders, especially depression [4,48–50]. First, let us consider the role of serotonin in clinical depression. In many patients, one effective treatment involves the use of a selective serotonin reuptake inhibitor (SSRI), which prolongs the action of serotonin at target sites. Data from animals suggests that serotonin release is involved in (learned) behavioral inhibition [50–53], associated with the prediction of aversive outcomes [54,55]. Computational modeling inspired by these data suggests that serotonin's role in behavioral inhibition may reflect a Pavlovian effect: subjects do not have to learn explicitly what (not) to do in the face of possible future trouble. This effect could be called the 'serotonergic crutch'. Problems with the operation of this crutch can lead to behavior in which poor choices are made because they have not been learned to be inappropriate. In this framework, punishments are experienced or imagined even if the choices concern internal trains of thought rather than external events [50].

Restoration of the crutch is considered to improve matters again. The logic here is that the more an individual's behavior is determined in a Pavlovian manner, the more devastating is the likely consequence of any problem with the serotonergic crutch. This is an account of vulnerability (analogous to the incentive sensitization theory of drug addiction; see [56–58]).

Conversely, model-based RL has been used to capture another feature of some forms of anxiety and depression: learned helplessness [59–61]. Animals can be made helpless when provided with uncontrollable rewards as well as uncontrollable punishments [62] and, thus, learning that their actions do not consistently predict outcomes. In these experiments, one way to demonstrate the onset of learned helplessness is to show that the animals do not explore or try to escape when placed in new environments (e.g. [63]).

A natural computational account is to treat the helplessness training in the first part of learned helplessness as inducing a prior probability distribution over possible future environments, indicating that the animal can expect to have little influence over its fate, that is, little controllability. This hypothesis is based on the expectation that related environments have similar properties. Exploration in a new environment is only worthwhile only if it is expected that good outcomes can be reliably achieved given appropriate actions. Thus, a prior belief implying that the environment is unlikely to afford substantial controllability will discourage exploration. Prior distributions are under active examination in Bayesian approaches to cognitive science and are offering substantial explanations for a broad range of developmental and adult behaviors [64]. Only model-based RL is capable of incorporating such rich priors, even though model-free control can be induced to behave in similar ways by simpler mechanisms [65,66]. This computational interpretation of uncontrollability provides a new way to understand the role that environments can play in etiology. It also provides a way of formalizing the complex interaction between model-based, model-free and Pavlovian systems, when not only might one controller directly influence the training signal of the other controllers [67] but also the very experience other controllers require in order to learn their own predictions or courses of action.

Using games to phenotype autism spectrum disorder

A defining feature of human cognition is the capacity to model and understand the intentions (and emotions) of other humans. This extends to an ability to forecast into the near-term future, for example, how someone else will feel should they experience a consequence of an action that we might take. Sophisticated capacities such as these lie at the heart of our ability to cooperate, compete and communicate with others. One of the defining features of autism spectrum disorder (ASD) is a diminished capacity for socio-emotional reciprocity – the social back-and-forth engagement associated with all human interaction [1,68–70]. Recent modeling and neuroimaging work has used two-agent interactions, typically in the form of some game, to parameterize and probe this social give-and-take [71–81]. This work, along with other efforts [82–85], has collectively launched a computational neuroscience perspective on inter-personal exchange – a first step toward identifying computational phenotypes in human interactions that are underwritten by both behavioral and neural responses.

Game theory is the study of mathematical models of interacting rational agents. It is used in many domains and in recent years has been increasingly applied to common behavioral interactions in humans. Two game-theoretic approaches have recently been used to probe ASD and other psychopathological populations directly: the stag hunt game and the multi-round trust game (Figure 2). Although the behavioral probes are different, the two

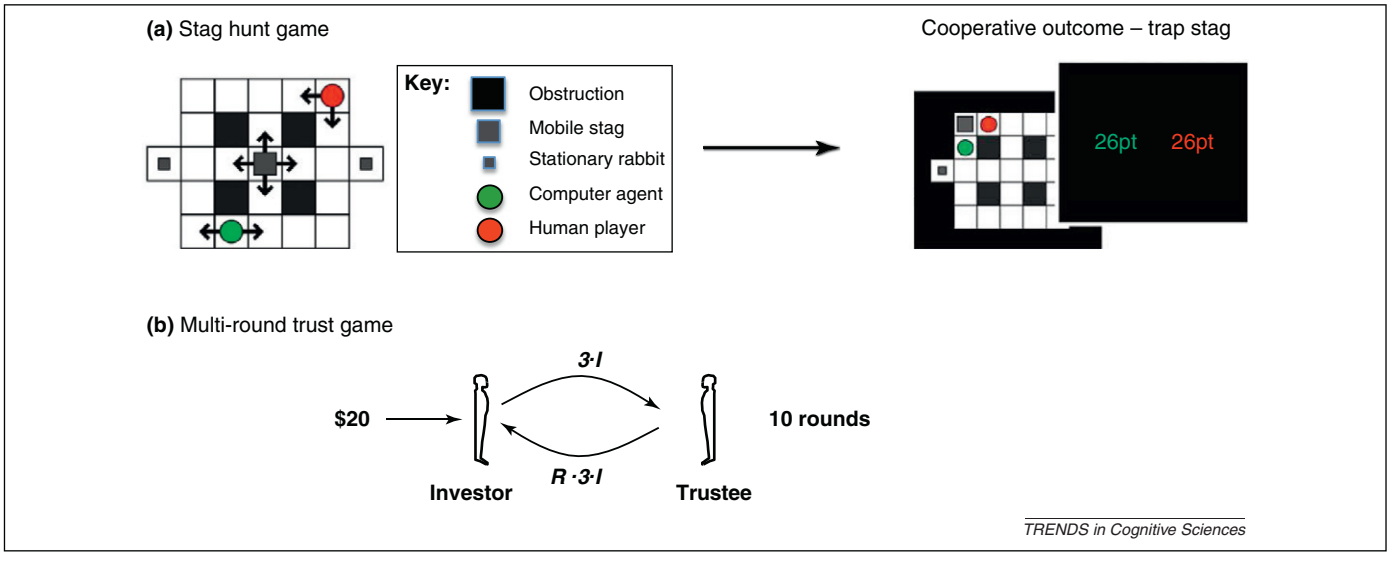

Figure 2. Economic games requiring theory-of-mind modeling. (a) The stag-hunt game. A two-player game, where players can cooperate to hunt and trap high yield stags or act alone and hunt low yield rabbits. A human subject (red circle) plays the game with a computer agent partner (green circle) to acquire either a (high value) mobile stag (larger gray square) or a (low value) rabbit sitting at a fixed position (smaller gray square). The hunters can catch a rabbit simply by moving onto its fixed location or they can catch a stag together by cooperatively trapping it somewhere on the open grid. **(b)** The multi-round trust game. A game of reciprocation that lasts 10 rounds. In each round, the proposer (the investor) is engaged with 20 monetary units. The investor any fraction of this (I) to the responder (the trustee). On the way to the trustee the amount triples and the trustee can repay any fraction R of the tripled amount. Players who think through the impact of their actions on their partner make more money than those who do not [73,78].

games share the feature that it is advantageous for a human player to make inferences about their partner's likely mental state during the game. These inferences are recursive: my model of you incorporates my model of your model of me, and so on. Both approaches have built computational models around this central idea of recursion [78,79], thereby furnishing component computations required for healthy human exchange.

Yoshida *et al.* [79–81] used the stag hunt game (Figure 2a) to probe mental state inferences in ASD versus control subjects. The stag hunt game is a classic two-player game (in this case involving a human player and a computer agent), where players can cooperate to hunt and acquire high yield stags, or act alone and hunt low yield rabbits. The model developed by Yoshida *et al.* [79] used the human player's observed behavior to estimate the sophistication level (depth of recursion) of their inference about the computer agent's beliefs (theory of mind). This estimate is necessary if the human player is to cooperate successfully with the computer agent – the human must believe that the agent believes that the human will also cooperate, and so on.

Behavioral results that exploited this model pointed to a higher probability for a theory-of-mind model (versus fixed-strategy) for control subjects. The opposite was true for ASD subjects (~78% probability for fixed-strategy). However, as one might expect, there was heterogeneity in these estimates, with some ASD subjects (n=5) displaying higher probability for the theory-of-mind model compared to a fixed-strategy model (n=12). Intriguingly, ASD subjects with a higher probability for a fixed-strategy model showed higher ratings on two ASD rating scales (ADI-R, ASDI). These results are preliminary and the sample of ASD subjects small. However, the crucial point is that the model allows for a principled parameterization of important cognitive components (e.g., depth of recursion in modeling one's partner). The use of such a model provides a way to formalize the cognitive components of ASD in computational terms. By collecting much more normative data, this type of approach could serve to differentiate ASD along these newly defined computational dimensions to improve diagnosis, guide other modes of investigation and help tailor treatments.

The multi-round trust game has also been used to probe a range of psychopathologic populations including ASD and borderline personality disorder [2,75,77]. The game is a sequential fairness game involving reciprocation, where performance is determined by whether players think through the impact of their actions on their partner. In the game, a proposer (called the investor) is endowed with $20 and chooses to send some fraction I to their partner. This fraction is tripled (to 3*I) on the way to the responder (called the trustee), who then chooses to send back some fraction of the tripled amount. Subjects play 10 rounds and know this beforehand. Cooperation earns both players the most money. Even when playing with an anonymous partner, investors do send money, a fact that challenges rational agent accounts of such exchanges. One way to conceptualize this willingness to send money was proposed by Fehr and Schmidt [86], who suggested that in such a social setting a player's utility for money depends on the fairness of the split across the two players. Based on this model of fair exchange between humans, Ray and colleagues developed a Bayesian model of how one player 'mentalizes' the impact of their actions (money split with partner) on their partner [78]. The key feature is for each player to observe monetary exchanges with their partner and estimate in a Bayesian manner the 'fairness type' of their partner, that is, the degree to which the partner is sensitive to an inequitable split.

This model was able to 'type' players reliably from 8 rounds of monetary exchange in the game. These types can be used to seek type-specific (fairness sensitivity) neural correlates. More importantly, the model can be used to phenotype individuals according to computational parameters important in this simple game-theoretic model of human exchange. This is an important new possibility. Using this same game, Koshelev and colleagues showed that healthy investors playing with a range of psychopathological groups in the trustee role can be clustered in a manner that reflects that type of psychopathology acting as the trustee [87]. This model used a Bayesian clustering approach to observations of the healthy investors' behavior as induced by interactions with different psychopathology groups. These preliminary results suggest that parameters extracted from staged (normative) game-theoretic exchanges could be used profitably as a new phenotyping tool for humans, where the phenotypes are defined by computational parameters extracted using models.

Computational phenotyping of human cognitive function

Computational models of human mental function present more general possibilities for producing new and useful human phenotypes. These phenotypes can then structure the search for genetic and neural contributions to healthy and diseased cognition. We do not expect such an approach to supplant current descriptive nosologies; instead, they will be an adjunct, where the nature of the computational characterization offers a new lexicon for understanding mental function in humans. Moreover, this approach can start with humans, define a computational phenotype, seek neural and genetic correlates of this phenotype and then turn to animal models for deeper biological study.

Under the restricted decision-making landscape that we have painted, RL models provide a natural example of a type of computational model that could be used in such phenotyping. Moreover, we sketched briefly how game-theoretic probes also allow for new forms of computational modeling and hence new ways to computationally phenotype humans. Through their built-in principles of operation and notions of optimal performance, RL models provide constraints that help bridge the aforementioned gap between molecular and behavioral levels of description. However, the behavioral underpinning of these models is extremely shallow at present, especially in human subjects. As suggested by the examples above, the estimation and use of computational variables, such as these, will require new kinds of behavioral probes, combined with an ever-evolving capacity to make neural measurements in healthy human brains. Not only is better phenotyping

through the development of new probes needed, but also unprecedented levels of phenotyping of cognitive function. Many of the best ideas about mental performance and function derive primarily from studies in other species. While these animal models have been strikingly successful at uncovering the biology underlying learning, memory and behavioral choice, the human behavioral 'software' is likely to be significantly different in important ways that the probes will need to capture. Large-scale computational phenotyping will require radical levels of openness across scientific disciplines and successful models for data exchange and data sharing.

Concluding remarks

If the computational approaches we have outlined turn out to be effective in psychiatry, then what might one expect? The large-scale behavioral phenotyping project sketched above involves substantial aspects of data analysis and computational modeling. The aim of the data analysis will be to link precise elements of the models to measurable aspects of behavior and to molecular and neural substrates that can be independently measured. A strong likelihood here is that the models will offer a set of categories for dysfunction that are related to, but different from, existing notions of disease and this will lead to a need for translation.

Although we did not focus on them here, there are also implications for mathematical modeling. A simulation-based account of measurable brain dynamics, anatomical pathways and brain regions could be expected, equipped with visualization and analysis methods to help make sense of the output. The ultimate hope is for a detailed, multi-level, model that allows prediction of the effects of malfunctions and manipulations. However, making this sufficiently accurate at the scales that matter for cognition and behavior is a long way off. One critical, though as yet unproven, possibility is that a computational understanding will provide its own kind of short-circuit, with, for instance, rules of self-organization of neural elements based on achieving particular computational endpoints, thereby removing the requirement for detailed specification.

Finally, the most pressing requirement is for training. Broad and deep skills across cognitive neuroscience, computational neuroscience, cellular and molecular neuroscience, pharmacology, neurology, and psychiatry itself, in addition to computer science and engineering, are required for the emergence of the richly interdisciplinary field of computational psychiatry. Optimistically, how to achieve this may become clearer as thoughts mature about restructuring education to achieve breadth across the brain-related clinical disciplines of neurology and psychiatry [88].

References

1 American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*, American Psychiatric Association
2 Kishida, K.T. *et al.* (2010) Neuroeconomic approaches to mental disorders. *Neuron* 67, 543–554
3 Maia, T. and Frank, M.J. (2011) From reinforcement learning models to psychiatric and neurological disorders. *Nat. Neurosci.* 14, 154–162
4 Huys, Q.J.M. *et al.* (2011) Are computational models useful for psychiatry? *Neur. Netw.* 24, 544–551
5 Hodgkin, A.L. and Huxley, A.F. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544
6 Dayan, P. and Abbott, L.F. (2001) *Theoretical Neuroscience*, MIT Press
7 Marreiros, A.C. *et al.* (2009) Population dynamics under the Laplace assumption. *Neuroimage* 44, 701–714
8 Tretter, F. *et al.* (2011) Affective disorders as complex dynamic diseases: a perspective from systems biology. *Pharmacopsychiatry* 44 (Suppl. 1), S2–S8
9 Moran, R.J. *et al.* (2011) Alterations in brain connectivity underlying Beta oscillations in parkinsonism. *PLoS Comput. Biol.* 7, e1002124
10 Wood, M.D. and Wren, P.B. (2008) Serotonin-dopamine interactions: implications for the design of novel therapeutic agents for psychiatric disorders. *Prog. Brain Res.* 172, 213–230
11 Friston, K.J. and Dolan, R.J. (2010) Computational and dynamic models in neuroimaging. *Neuroimage* 52, 752–765
12 Honey, C.J. *et al.* (2007) Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10240–10245
13 Turing, A.M. (1950) Computing machinery and intelligence. *Mind* 59, 433–460
14 Gödel, K. (1931) Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme, I. *Monatshefte für Mathematik und Physik* 38, 173–198
15 Turing, A.M. (1936) On computable numbers, with an application to the entscheidungsproblem. *Proc. Lond. Math. Soc.* 42, 230–265
16 Church, A. (1936) An unsolvable problem of elementary number theory. *Am. J. Math.* 58, 345–363
17 Nagel, E. and Newman, J.R. (1958) *Gödel's Proof*, New York University Press
18 Garey, M.R. and Johnson, D.S. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman
19 Friston, K.J. (2010) The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138
20 Kamil, A.C. *et al.* (1987) *Foraging Behavior*, Plenum Press
21 Uhlhaas, P.J. and Singer, W. (2011) The development of neural synchrony and large-scale cortical networks during adolescence: relevance for the pathophysiology of schizophrenia and neurodevelopmental hypothesis. *Schizophr Bull.* 37, 514–523
22 Servan-Schreiber, D. *et al.* (1990) A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. *Science* 249, 892–895
23 Montague, P.R. *et al.* (1994) Foraging in an uncertain environment using predictive Hebbian learning. *Adv. Neural Inform. Proc. Sys.* 6, 598–605
24 Montague, P.R. *et al.* (1995) Bee foraging in uncertain environments using predictive Hebbian learning. *Nature* 377, 725–728
25 Montague, P.R. *et al.* (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* 16, 1936–1947
26 Montague, P.R. *et al.* (2004) Computational roles for dopamine in behavioral control. *Nature* 431, 760–767
27 McCulloch, W. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133
28 Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408
29 McClelland, J. and Rumelhart, D. (1989) *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*, MIT Press
30 O'Reilly, R. and Munakata, Y. (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*, MIT Press
31 Hopfield, J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558
32 Hopfield, J.J. (1984) Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. U.S.A.* 81, 3088–3092
33 Hopfield, J.J. and Tank, D.W. (1986) Computing with neural circuits: a model. *Science* 233, 625–633

34 Hinton, G.E. and Sejnowski, T.J. (1983) Optimal perceptual inference, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington DC. pp. 448–453
35 Cohen, J.D. and Servan-Schreiber, D. (1992) Context, cortex and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychol. Rev.* 99, 45–77
36 Cohen, J.D. and Servan-Schreiber, D. (1993) A theory of dopamine function and cognitive deficits in schizophrenia. *Schizophr. Bull.* 19, 85–104
37 Cohen, J.D. *et al.* (1996) A computational approach to prefrontal cortex, cognitive control, and schizophrenia: Recent developments and current challenges. *Phil. Trans. R. Soc. Lond. B: Biol. Sci.* 351, 1515–1527
38 Braver, T.S. *et al.* (1999) Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. *Biol. Psychiatry* 46, 312–328
39 Carter, C.S. *et al.* (1998) Anterior cingulate cortex, error detection, and the on line monitoring of performance. *Science* 280, 747–749
40 Carter, C.S. *et al.* (2001) Anterior cingulate cortex and impaired self-monitoring of performance in patients with schizophrenia: an event-related fMRI study. *Am. J. Psychiatry* 1423–1428
41 Frank, M.J. *et al.* (2004) By carrot or by stick: cognitive reinforcement learning in Parkinsonism. *Science* 306, 1940–1943
42 O'Reilly, R.C. (2006) Biologically-based computational models of high-level cognition. *Science* 314, 91–94
43 O'Reilly, R.C. and Frank, M.J. (2006) Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.* 18, 283–328
44 Hazy, T.E. *et al.* (2007) Toward an executive without a homunculus: bomputational models of the prefrontal cortex/basal ganglia system. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 362, 1601–1613
45 Smith, A.J. *et al.* (2007) Linking animal models of psychosis to computational models of dopamine function. *Neuropsychopharmacology* 32, 54–66
46 Corlett, P.R. *et al.* (2011) Glutamatergic model psychoses: prediction error, learning, and inference. *Neuropsychopharmacology* 36, 294–315
47 Redish, A.D. *et al.* (2008) A unified framework for addiction: vulnerabilities in the decision process. *Behav. Brain Sci.* 31, 415–437 (discussion pp. 437-487)
48 Kumar, P. *et al.* (2009) Abnormal temporal difference reward-learning signals in major depression. *Brain* 131, 2084–2093
49 Gradin, V.B. *et al.* (2011) Expected value and prediction error abnormalities in depression and schizophrenia. *Brain* DOI: 10.1093/brain/awr059
50 Dayan, P. and Huys, Q.J.M. (2009) Serotonin in affective control. *Annu. Rev. Neurosci.* 32, 95–126
51 Soubrie, P. (1986) Reconciling the role of central serotonin neurons in human and animal behavior. *Behav. Brain Res.* 9, 319–364
52 Boureau, Y-L. and Dayan, P. (2011) Opponency revisited: competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology* DOI: 10.1038/npp.2010.151
53 Cools, R. *et al.* (2008) Acute tryptophan depletion in healthy volunteers enhances punishment prediction but does not affect reward prediction. *Neuropsychopharmacology* 33, 2291–2299
54 Deakin, J.F.W. (1983) Roles of brain serotonergic neurons in escape, avoidance and other behaviors. *J. Psychopharmacol.* 43, 563–577
55 Deakin, J.F.W. and Graeff, F.G. (1991) 5-HT and mechanisms of defense. *J. Psychopharmacol.* 5, 305–316
56 Robinson, T.E. and Berridge, K.C. (2000) The psychology and neurobiology of addiction: an incentive-sensitization view. *Addiction* 95, s91–s117
57 Robinson, T.E. and Berridge, K.C. (2008) The incentive sensitization theory of addiction: some current issues. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 363, 3137–3146
58 Flagel, S.B. *et al.* (2009) Individual differences in the attribution of incentive salience to reward-related cues: implications for addiction. *Neuropharmacology* 56 (Supp. 1), 139–148
59 Seligman, M.E.P. (1975) *Helplessness on Depression, Development and Death*, W.H. Freeman & Co.
60 Maier, S. and Seligman, M. (1976) Learned helplessness: Theory and evidence. *J. Exp. Psychol. Gen.* 105, 3–46
61 Miller, W.R. and Seligman, M.E. (1975) Depression and learned helplessness in man. *J. Abnorm. Psychol.* 84, 228–238
62 Goodkin, F. (1976) Rats learn the relationship between responding and environmental events: An expansion of the learned helplessness hypothesis. *Learn. Motiv.* 7, 382–393
63 Maier, S.F. and Watkins, L.R. (2005) Stressor controllability and learned helplessness: The roles of the dorsal raphe nucleus, serotonin, and corticotropin-releasing factor. *Neurosci. Biobehav. Rev.* 29, 829–841
64 Tenenbaum, J.B. *et al.* (2011) How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285
65 Sutton, R.S. (1990) Integrated architectures for learning, planning, and reacting based on approximating dynamic programming, In *Proceedings of the 7th International Conference on Machine Learning*, Morgan Kaufmann. pp. 216–224
66 Kakade, S. and Dayan, P. (2002) Dopamine: generalization and bonuses. *Neur. Netw.* 15, 549–559
67 Daw, N.D. *et al.* (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69, 1204–1215
68 Baron-Cohen, S. (2001) Theory of mind and autism: a review. *Int. Rev. Res. Ment. Retard.* 23, 169–184
69 Frith, C.D. and Frith, U. (1999) Interacting minds – a biological basis. *Science* 286, 692–1695
70 Gallagher, H.L. and Frith, C.D. (2003) Functional imaging of 'theory of mind'. *Trends Cogn. Sci.* 7, 77–83
71 Rilling, J.K. *et al.* (2002) A neural basis for social cooperation. *Neuron* 35, 395–405
72 Rilling, J.K. and Sanfey, A.G. (2011) The neuroscience of social decision-making. *Annu. Rev. Psychol.* 62, 23–48
73 King-Casas, B. *et al.* (2005) Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78–83
74 Tomlin, D. *et al.* (2006) Agent-specific responses in cingulate cortex during economic exchanges. *Science* 312, 1047–1050
75 Chiu, P.H. *et al.* (2008) Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism. *Neuron* 57, 463–473
76 Hampton, A.N. *et al.* (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc. Natl. Acad. Sci. U.S.A.* 105, 6741–6746
77 King-Casas, B. *et al.* (2008) The rupture and repair of cooperation in borderline personality disorder. *Science* 321, 806–810
78 Ray, D. *et al.* (2008) Bayesian model of behavior in economic games. *Adv. Neural Inform. Proc. Sys.* 21, 1345–1353
79 Yoshida, W. *et al.* (2008) Game theory of mind. *PLoS Comput. Biol.* 4, DOI: 10.1371/journal.pcbi.1000254
80 Yoshida, W. *et al.* (2010) Neural mechanisms of belief inference during cooperative games. *J. Neurosci.* 30, 10744–10751
81 Yoshida, W. *et al.* (2010) Cooperation and heterogeneity of the autistic mind. *J. Neurosci.* 30, 8815–8818
82 Bhatt, M. and Camerer, C. (2005) Self-referential thinking and equilibrium as states of minds in games: FMRI evidence. *Games Econ. Behav.* 52, 424–459
83 Coricelli, G. and Nagel, R. (2009) Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9163–9168
84 Bhatt, M.A. *et al.* (2010) Neural signatures of strategic types in a two-person bargaining game. *Proc. Natl. Acad. Sci. U.S.A.* 107, 19720–19725
85 Lee, D. (2008) Game theory and neural basis of social decision making. *Nat. Neurosci.* 11, 404–409
86 Fehr, E. and Schmidt, K.M. (1999) A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868
87 Koshelev, M. *et al.* (2010) Biosensor approach to psychopathology classification. *PLoS Comput. Biol.* 6, e1000966
88 Insel, T.R. and Wang, P.S. (2010) Rethinking mental illness. *JAMA* 303, 1970–1971
89 Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning*, MIT Press
90 Daw, N.D. and Doya, K. (2006) The computational neurobiology of learning and reward. *Curr. Opin. Neurobiol.* 16, 199–204
91 Dayan, P. and Daw, N.D. (2008) Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci.* 8, 429–453
92 Tolman, E.C. (1948) Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208
93 Doya, K. (1999) What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex. *Neural Netw.* 12, 961–974

94 Dickinson, A. and Balleine, B. (2002) The role of learning in motivation, In *Stevens' Handbook of Experimental Psychology, vol. 3: Learning, Motivation and Emotion* (3rd ed.) (Gallistel, C.R., ed.), pp. 497–533, Wiley

95 Killcross, S. and Coutureau, E. (2003) Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb. Cortex* 13, 400–408

96 Daw, N.D. *et al.* (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711

97 Glascher, J. *et al.* (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66, 585–595