# 2

# Breakdowns and Failure Modes

## An Engineer's View

A. David Redish and Joshua A. Gordon

**Abstract**

Psychiatry faces a number of challenges due largely to the complexity of the relationship between mind and brain. Starting from the now well-justified assumption that the mind is instantiated in the physical substrate of the brain, understanding this relationship is going to be critical to any understanding of function and dysfunction. Key to that translation from physical substrate to mental function and dysfunction is the computational perspective: it provides a way of translating knowledge and understanding between levels of analysis (Churchland and Sejnowski 1994). Importantly, the computational perspective enables translation to both identify emergent properties (e.g., how a molecular change in a receptor affects behavior) and consequential properties (e.g., how an external sociological trauma can lead to circuit changes in neural processing). Given that psychiatry is about treating harmful dysfunction interacting across many levels (from subcellular to sociological), this chapter argues that the computational perspective is fundamental to understanding the relationship between mind and brain, and thus offers a new perspective on psychiatry.

### The Computational Perspective

Fundamentally, the computational perspective is about how information is processed within neural circuits; it uses formal methods to identify how inputs and recurrent processing combine to create outputs. By being formal, the computational perspective enables an explanation and understanding of neuropsychology in its elemental basis so that we can identify measurable changes and determine where breakdowns occur. This perspective allows us to define *computational psychiatry* as a methodology using formal computational perspectives to address psychiatric dysfunction. It is only with computational explanations of function that we can begin to identify how physiological, sociological, and other changes can lead to *dys*function.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

The computational perspective hypothesizes that the role of brains is to perform computations to improve behavior. For instance, maintaining thermodynamic equilibrium is a complex computational process (Goldstein and McEwen 2002), as is sensory recognition for improved motor control (Llinas 2001), escape from predators or the identification of prey (Eaton 1984), and social prediction for interaction with conspecifics (Cheney and Seyfarth 1990). To understand how the brain computes these mental processes, we need to understand the mechanism of the computational process. While the definition may sound tautological, the key to the computational perspective, whether in computational neuroscience or computational psychiatry, is that it requires explanations to be specified in a formalism that forces a more complete story and often reveals obscure consequences. The process by which pathologies in physiological processes engender pathologies in psychological processes is often not obvious: complex consequences can only be derived from computational models and formal analyses.

When talking about a computational perspective, it is important to make clear what it is not. Although computer models can play important roles in computational psychiatry, it is possible to construct computational formalisms that provide explanations without an explicit computer model of pathopsychology. For instance, Kurth-Nelson et al. (this volume) describe examples of formalisms of families of models that are all susceptible to specific pathophysiologies. Similarly, computational psychiatry is more than applying computer algorithms to large data sets, such as clustering genetics or behavioral distributions or what is known as "big data" (Pevsner 2005; Schadt et al. 2010; Mayer-Schönberger and Cukier 2013). Neither computer models nor algorithmic clustering of large data sets provides understanding (although both can guide and test development of theories). For instance, Flagel et al. (this volume) provide a novel formalism for nosology which implies a clear use for big data in deriving consequences of diagnosis and treatment, but emphasizes the importance of the computational perspective to provide the underlying latent constructs.

We can define a theory as an explanation of how an observation arises from a lower-level phenomenon, such as how Parkinsonian movement disorders could arise from circuit changes derived from changes in dopaminergic tone (Albin et al. 1989). Models can then be used to test these theoretical hypotheses. For instance, the theory that Parkinsonian movement disorders arise from depletion of dopaminergic tone implies that one can create an animal model of Parkinson disease by depleting dopamine from an otherwise normal animal subject (Langston et al. 1984; Deumens et al. 2002). This animal model is really creating an analogous situation to that of Parkinson disease, which can then be explored. Similarly, a computational model can be used as an analogous situation which can then be explored. For instance, in a recent paper, Schroll et al. (2015) examined a detailed computational model of three theories of the progress of Huntington disease and found that only a progressive degeneration of medium spiny neurons in the direct and indirect pathways provides

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

compatible behavioral deficits to those seen in real patients. However, just as the animal model embodies a theory of Parkinsonian mechanism, but is not the theory of Parkinson disease, neither does a computational model deliver a theory of Huntington disease. Rather, these models are only interpretable when taken from a theoretical (computational) perspective about the computations being performed by the direct and indirect pathways of the basal ganglia (Albin et al. 1989; Kravitz et al. 2012).

Importantly, it is not necessary to provide an explanation down to the cellular, subcellular, or molecular level; the explanation has to be focused at the appropriate explanatory level. For instance, it is well established that hippocampal cells encode information about the spatial location of an animal ("place cells"; O'Keefe and Dostrovsky 1971; O'Keefe 2015). What mechanisms make these cells fire in their given locations is an interesting scientific question. However, if we want to know how a complex firing pattern in hippocampal place cells drives behavior, then it is not necessary to know why the place cells fire in that complex pattern, only that they do. An appropriate theory would start from the statement that hippocampal place cells show a complex spatially related firing pattern and use that to explain how changing those firing patterns changes navigational processes.

The usefulness of computational models should not be underestimated. Computational models force researchers to develop precise, explicitly specified, falsifiable hypotheses. Although it is the theoretical statements (neural mechanism *X* implies behavioral change *Y*, behavioral incident *Z* creates neural mechanism *X*) that will necessarily drive understanding and treatment, these derivations are rarely straightforward and rarely simple. If you think that neural mechanism *X* can drive behavior *Y*, then it should be possible to build a model of *X* that can perform behavior *Y*. There are many cases in neuroscience where people have thought that two effects were incompatible, only to find them compatible after a computational model was built (e.g., space and memory in hippocampus; Redish 1999). Biology is complex and neuroscience particularly so. It is dangerous to simply infer backward from symptoms to dysfunction. Flagel et al. (this volume) and Moran et al. (this volume) suggest a new computational formalism to provide a more nuanced inference process through computational mechanisms to connect symptoms, dysfunction, diagnosis, and treatment.

## What Does the Computational Perspective Provide?

Computational perspectives bring two things to the table that we believe fundamentally change psychiatry. First, a computational perspective promises to ask different questions about patients than traditional clinical perspectives. These novel questions can guide diagnosis and treatment by getting at fundamental psychological and neural processes which cut across symptomatic and

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

diagnostic boundaries. Computational perspectives suggest that the fundamental question we should be asking is: *What is different about how this patient processes information about the world (including the patient's self)?*

Second, computation is a way of addressing how mechanisms translate between levels, such as how a change in neural structure can lead to a change in behavior, or how an external incident can lead to changes in neural processing. For instance, a deficit in a certain ion channel in a specific neural structure changes the excitatory-inhibitory balance within that neural structure, and can produce behavioral changes observable at a macro level as epilepsy (Soltesz and Staley 2008). Similarly, externally induced (sociological) stress produces changes in hormone levels, neural circuits, and thus the computations a patient makes about interactions with the world (Payne et al. 2007).

These computational perspectives can help psychiatry incorporate increasingly more biological mechanisms into its categorization and treatment. However, as argued in this book, the computational perspective goes beyond connecting biology to psychiatry; psychiatric disorders themselves should be couched in terms of disorders of information processing and computations.

## The Failure Mode Hypothesis

Psychiatry starts from the concept of *harmful dysfunction*; that is, there is some underlying dysfunction in the system that is serious enough to warrant intervention and treatment (Wakefield 1992a, 2007; Flagel et al., this volume). Defining psychiatry from dysfunction implies that one must first understand *function* before one can see how it has become disrupted.

We will not assume that the starting dysfunction always proceeds from brain to behavior. For instance, many theories of gambling and addiction suggest that dysfunction arises from interactions between functional neural processes and dysfunctional external situations for which humans have not evolved to accommodate (Wagenaar 1988; Redish et al. 2007; Schüll 2012). While altering the brain changes the mind, the physical nature of the mind implies that altering the mind also changes the brain. Thus, translating across levels (whether from brain to mind or mind to brain) requires an understanding of the computational mechanics of that system, and how those computational mechanics change when underlying structures change. This computational perspective suggests that we can take the engineer's view and ask new questions about how the system can break down.

Applying an engineer's analysis to a specific dysfunctional or misbehaving system means trying to discover what the changes are in the system that have created the problematic behavior. These potential ways that a system can break down are known as "failure modes." Colloquially, we can think of this as: *Where are the weak links? Where and how does this system typically break?*

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

Importantly, errors (failure modes) can exist at many levels. Huntington disease is a genetic abnormality, a CAG repeat in the Huntingtin gene (Kremer et al. 1994). Parkinsonianism occurs from loss of dopaminergic function, which can arise from genetic dysfunction (Gasser 2009; Shulman et al. 2011) or an external toxin, such as MPTP (Langston et al. 1983; Langston and Palfreman 2013). Physical trauma (e.g., traumatic brain injury) creates an abnormality in the physical structure of the network. Mental trauma (e.g., from prolonged solitary confinement) creates changes in underlying structure leading to a change in the function of neural circuits (Grassian 1983). Drug addiction is an interaction between external causes (the drugs) and internal effects (neural susceptibility). All of these effects, however, fundamentally affect behavior by altering the brain's computation. To understand how these effects occur, we have to understand that computation and how it becomes altered.

From a treatment perspective, the engineering analysis attempts to find the levers of control: *Where are the points that can provide the optimal means of changing the system back into function?* Just as there can be errors (failure modes) at many levels, treatment can occur at many levels: pharmacological manipulation (Schatzberg and Nemeroff 1995), circuitry manipulations (Obeso and Guridi 2001; Mayberg et al. 2005), physical or mental training (Bickel et al. 2011), or even reinterpretations engendered by cognitive reprocessing (Ainslie 1992, 2001; Kurth-Nelson and Redish 2012) or changes in social interactions (Heyman 2009; Petry 2012). All treatments change the underlying physical and mental structure and thus the computational processing. Understanding how the computational process changes in the face of treatment is an important step in understanding when and how treatment should be used.

## Computational Perspectives throughout the Components of Processing

We can apply computational perspectives to all aspects of a person's interaction with the world. New computational perspectives have changed how we understand many processes, including decision making, memory, perception, and action.

For instance, computational and theoretical neuroscientists working on decision making have now garnered considerable evidence that there are several different action-selection systems, each of which processes information about the world differently (Redish 2013). Errors can occur within each of these different systems as well as in the interaction between these systems. Failure modes of the decision-making system will arise with drug addiction (Redish et al. 2008), emotional disorders, such as anxiety and depression (Huys 2007; Rangel et al. 2008) as well as motivational disorders, such as obsessive-compulsive disorder (Pitman 1987; Maia and McClelland 2012). Different treatments will be needed for different failure modes (Rangel et al. 2008; Redish et

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

al. 2008; Redish 2013). On the other hand, it may be possible to use an intact action-selection system to counter a failure mode within one of the other systems, either through training (Bickel et al. 2011, 2015) or through changing the situations the subject is experiencing (Heyman 2009; Kurth-Nelson and Redish 2012; Petry 2012).

Although there has been a tremendous amount of work in the relationship between decision making and psychiatry (Huys 2007; Rangel et al. 2008; Redish et al. 2008; Maia and Frank 2011; Montague et al. 2012; van der Meer et al. 2012b; Redish 2013), a similar computational perspective can be applied to other neural components. For instance, one can derive computational explanations for disordered thinking in schizophrenia (Seamans and Yang 2004; Tanaka 2006; Durstewitz and Seamans 2008). Computational perspectives on perception as signal detection (Tougaard 2002), information derivation (Poggio and Bizzi 2004; Serre et al. 2007), or situation recognition and categorization (Redish et al. 2007; Gershman and Niv 2010) can be used to explain disorders of perception, such as in hallucinations (Bressloff et al. 2002), migraine auras (Reggia and Montgomery 1996; Dahlem and Chronicle 2004), or an inability to recognize social cues (Dapretto et al. 2006; Singer 2008).

## What Can Computational Perspectives Provide to Psychiatry?

The companion introductory paper (Gordon and Redish, this volume) describes psychiatry as facing three current challenges and notes three current promises being incorporated into psychiatry. We believe that the computational viewpoint can provide help with these challenges and provides novel perspectives on these promises.

### Challenge 1: Nosology

Psychiatry has long noted the difference between reliable and valid categories (McHugh and Slavney 1998). A reliable category is one where membership can be reliably assigned. In contrast, a valid category is one that reflects an underlying similarity in process or outcome. As noted in the companion paper (Gordon and Redish, this volume), psychiatry has a high degree of inter-rater reliability that is similar to many medical disorders, but the efficacy of psychiatric treatments lags that of many other medical disorders. This seems to be because the categories (while reliable) do not conform to either biological or treatment boundaries.

In part, this lack of efficacy is because psychiatry has long found itself trapped between the necessity of using categorical constructs for diagnosis (e.g., ICD-10 or DSM-IV-TR) and more parsimonious dimensional constructs to explain behavior (Krueger 1999; Insel et al. 2010; see also chapters by First, MacDonald et al., and Friston, this volume). The computational perspective

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

can provide a novel solution to this category/dimension complexity by integrating both together.

Nosology has a number of goals ranging from communication (between clinicians, as well as to patients and their families) to guiding diagnosis and treatment (see chapters by First as well as Flagel et al., this volume). In part, categories have arisen because of their enhanced simplicity in identifying diagnosis (for insurance reimbursement purposes) and treatment. Given that treatment is in the end an action taken, treatment is a categorical decision (either you treat or you don't). However, the computational perspective provides a more nuanced perspective on this categorical decision.

Several chapters in this book lay out a new computational nosology based on Bayesian inference linking underlying dimensional constructs with categorical diagnoses and actions (Flagel et al., Moran et al., and Friston, this volume). In short, an underlying set of dimensional constructs (computational constructs) predicts observations through a well-defined (formal) mathematical definition known as Bayesian inference (for mathematical details, see chapters by Mathys and Friston, this volume.) This inference process allows inferences to proceed both from observations to constructs and from constructs to observations. In this formulation, observations can be measurements (such as from a biological or behavioral test), psychological instruments (such as answers on a questionnaire), or diagnoses.

Importantly, in this formulation, diagnoses are seen not as a direct reflection of the fundamental dysfunction, but rather as clinical observations that arise from underlying (computational) dysfunctions. For instance, the hypothesis that repeated drug use can arise from many potential failure modes in different neural systems suggests that the clinical identification of dependence is only a partial predictor of the potential underlying dysfunctions (Redish et al. 2008). Similarly, treatment is seen as changing the trajectory of the dimensional (computational) constructs over time and thus as changing the future observations (symptoms).

This new *computational nosology* relies on two aspects of the computational perspective: the multifarious nature of the relationship between the underlying dysfunctions and diagnoses, and the computational nature of the underlying constructs.

## Challenge 2: Biomarkers

The computational nosology hinted at in the preceding section and developed fully by Flagel et al., Moran et al., and Friston (this volume), imply a new perspective on biomarkers. The use of computation to drive nosology suggests that computational perspectives should provide new opportunities to identify differences along construct dimensions and thus to measure differences between categories. A classic example is the way that an EKG measures the dynamics of the heart electrophysiology, allowing separation of chest pain.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

Heart attack and indigestion both create chest pain, but require different treatments—having an EKG to measure heart function can be critical to identifying the appropriate treatment. One of the main advantages of the new nosology proposed in several chapters in this volume (see MacDonald et al., Flagel et al., Moran et al., and Friston) is that biomarkers don't necessarily have to define categorical diagnoses, as long as they can be used to drive predictions about prognosis and/or treatment response.

It is likely that computation itself will provide new (bio)markers for use throughout psychiatry. In fact, those markers do not have to be biological per se, but might instead be detectable through behavioral tests or even instrumental questionnaires. As noted above, the computational perspective changes the question of psychiatric function and dysfunction to "*What is different about how this patient processes information about itself or about the world?*" This suggests that measures of computational processing can be used to differentiate patients and treatment. For instance, smokers deal with counterfactual (could-have-been) rewards differently than nonsmokers (Chiu et al. 2008). Huntington patients are less able to compensate for force changes applied to motor activities (Smith et al. 2000). And drug-dependent users (on average) discount future outcomes at much faster rates than non-users (Kirby et al. 1999; Odum et al. 2002). However, within any drug-dependent population, some users do show normal discounting rates. Recently, an analysis of drug-addiction treatments found that the more successful treatments normalized the discounting rates of the subset of users who were discounting overly fast (Bickel et al. 2014). This tells us two things—first, that these treatment processes are not just selecting for the subset of users with normal discounting rates (so we can send both fast and slow discounting users to treatment), and second, that there is a relationship between treatment success and changes in discounting rate (which imply a potential marker for treatment success, at least in a subset of patients).

Computational perspectives can also be used to identify how biomarkers produce their effects and what effects they are likely to produce. For instance, changes in genetic underpinnings of dopamine receptor efficacy (in D1, D2, and the COMT variation) produce differences in the efficacy of learning strategies—genetic changes in D1 drive learning from positive rewards, whereas genetic changes in D2 drive learning from punishment signals, and the COMT variation affects ability to reverse responses (Frank et al. 2007a). From a computational understanding of dopamine's role in driving learning, the effects of Parkinson disease (decreasing dopaminergic tone, leading to lower signal-to-noise ratios between phasic bursts of dopamine and baseline levels, making it hard for neurons to recognize phasic increases in dopamine) and levodopa treatment (increasing dopaminergic tone, but reducing the depth of the drop in dopamine that occurs with undelivered reward, thus making it harder for neurons to recognize decreases in dopamine signals) can be predicted (Frank 2011; Moustafa and Gluck 2011).

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

These predictions have been confirmed experimentally (Frank et al. 2004; Shohamy et al. 2006; Kéri et al. 2010). Whether they can be used to guide treatment remains an open question.

In contrast, the evidence that hippocampal size is a biomarker for a vulnerability to posttraumatic stress disorder (PTSD) in the face of trauma (Gilbertson et al. 2002) does not explain why decreased hippocampal size predicts that vulnerability. Computational explanations of hippocampal dysfunction, such as inability to recognize context (Nadel and Jacobs 1996; Jacobs and Nadel 1998) or an inability to consolidate memories (Redish 2013), may provide some clues. Computational analyses of hippocampally dependent behaviors would likely predict that vulnerability as well, and may be easier to test potential patients behaviorally than with structural MRI scans. (Imagine, e.g., if every soldier was tested for a vulnerability to PTSD by a behavioral measure of hippocampal abilities before being sent into combat).

It will be important to determine which of these computational changes are trait effects (preexisting within the individual) and which are state effects (thus temporary effects due to the physical, mental, and pharmacological situation in which the individual finds itself). While it has long been hypothesized that discounting impulsivities could drive addiction (Ainslie 1992; Belin et al. 2008; Odum and Baumann. 2010), the fact that discounting changes normalize after treatment (Odum et al. 2002; Bickel et al. 2014) suggests that discounting differences with addicts may be more state effects than trait effects. Both state effects and trait effects can be useful biomarkers. At this point, most computational biomarkers have not been as thoroughly examined as the discounting of future options, but computational biomarkers have been proposed for other dysfunctions as well, such as personality disorders and sociopathy differences that can be tested through economic games such as the ultimatum, trustee, or dictator games (Kishida et al. 2010).

### Challenge 3: Treatment

Computation provides new perspectives not only on dysfunction (failure modes) but also on how the system can change (where the levers are). This means that computational analyses of treatment paradigms could provide better explanations for how those treatments are working, which can suggest new ways to improve them. Computational analyses should also provide better explanations for which dysfunctions will be ameliorated by treatments, which can suggest better assignments of patients to treatment.

As nosologies and biomarkers are improved, it should become easier to assign specific treatments to specific patients. In addition, we suspect that new treatments will be developed as the underlying computational dysfunctions are identified and as the available manipulations are identified. Computational analyses of treatment will suggest which subsets of patients (i.e., which failure modes) will be best ameliorated by a given treatment. At the extreme, this

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

leads to personalized medicine—identifying a specific battery of treatments optimized for the specific computational processes underlying a given person's mind and brain.

At this point, the computational implications of treatment are only just starting to be explored. For instance, Regier and Redish (2015) suggest that the treatment of contingency management, in which rewards are provided to addicts for staying clean of drugs, is unlikely to be working primarily through the basic economics of making drugs more expensive (taking drugs now loses the addict the alternate reward as well as the usual cost of the drugs), because a computational analysis showed that the alternate rewards are too small. Instead, they suggest that contingency management creates an explicit choice between two concrete rewards (small as one of them may be). Explicit choice ("Take drugs or get that gold star") tends to activate different decision-making systems that are dependent on different neural circuits from Go/No-Go decisions ("take drugs or don't"). While it is not known whether this computational explanation for contingency management is correct or not, this hypothesis suggests that contingency management success would depend on the neural circuits that drive explicit choice behaviors (such as prefrontal–hippocampal interactions). These circuits could be examined (e.g., through structural or functional imaging), and if impaired, could be improved, either through training or through pharmacological means. There is some evidence that prefrontal integrity protects against relapse after treatment (Camchong et al. 2014), and animal models have suggested that pharmacological interventions can change prefrontal integrity (Dalley et al. 2004). Because of the computational understanding of the processes that underlie explicit choice decision making, it is also possible to suggest ways to change the contingency management treatment itself. For instance, making delayed options more explicitly concrete activates the prefrontal–hippocampal interaction and makes people more willing to wait for delayed rewards (Peters and Büchel 2010). This suggests that very concrete options would be particularly effective in contingency management.

**Promise 1: Genetics**

As noted in the companion paper (Gordon and Redish, this volume), there is a strong genetic component to psychiatric disorders. However, the connection from genetics to behavior (and worse, to dysfunctional behavior) is long and arduous and depends on interactions with many other environmental and social components. It is rare that a single genetic abnormality translates to a behavioral disorder (such as Huntington disease), but even in those cases, the behavioral consequences can be quite complex and varied. In other disorders, hundreds of genetic markers have been found, which suggests that these markers are being transformed through some intervening substrate to generate the behavior.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

The computational perspective provides an access point to this intervening substrate: genetics change the physical nature of neural circuits, which changes how they compute. Thus, if we want to understand the role of genetics in psychiatric disorders, we need to use computation to connect genetics with neural circuits and neural circuits with behavior. Some computational progress has been made (e.g., in epilepsy) where many genetic abnormalities have the common effect of changing the balance between excitation and inhibition, which creates a mathematical instability and the potential for a sudden shift from one state (balance) to another (seizure) (Soltesz and Staley 2008).

**Promise 2: Circuits**

The translation through neural circuits suggests the possibility of identifying psychiatric disorders at the level of those circuits themselves. But just as one needs computation to translate from genetic changes to functional changes in neural circuits, one needs computation to translate from physical changes in neural circuits to behavioral changes. Computational neuroscience has been quite successful over the last several decades identifying how functional circuits create behavior, particularly in the context of normal function and constructed dysfunction.

For instance, it was computational perspectives that suggested a role for the hippocampus in spatial navigation (O'Keefe and Nadel 1978) and led to the development of the water maze (Morris 1981), including the subtle distinctions between variants of it (Eichenbaum et al. 1990; Day et al. 1999). Similarly, computational perspectives have been critical for an understanding of memory transformations and the development of schema and the transformation from episodic to semantic memory (O'Reilly and McClelland 1994; Redish and Touretzky 1997). Computational perspectives on multiple decision-making systems successfully predicted how different neural circuits process information about cues, actions, and rewards (van der Meer et al. 2010). It has long been possible to manipulate circuits, but new techniques can provide exquisite control of neural circuits at unprecedented cellular, connectivity, and temporal scales (Tye and Deisseroth 2012). With advances in both our computational understanding of neural circuits and newly available techniques to manipulate them neurophysiologically, it is now becoming possible to create explicit deficits hypothesized by computational theories and to test whether they produce the expected behavioral consequences. Whether that success can be translated into clinical practice, however, remains an open question.

**Promise 3: Personalized Medicine**

To tailor therapy to an individual patient, we need to understand both dysfunction and treatment at a deep enough level so that we can match treatment to dysfunction. We argue that computation is the path to that understanding.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

By understanding how physical dysfunctions create psychiatric disorders, we can identify the most appropriate probes to identify which dysfunctions exist within a given patient. By understanding how treatment changes the brain (and thus the mind), we can identify the most appropriate treatments (and the most appropriate variations on a given treatment) for a given patient. This is the promise of the computational perspective. But it raises the obvious question: *Are we there yet?*

## Open Questions

The first, and most important question, is whether the computational perspective is ready to take to the clinic yet. There are obviously large gaps in our knowledge of the computations being performed by neural systems as well as large gaps in our knowledge of how neural circuits (and subcircuit dynamics) create those computations. However, as we have seen throughout this chapter (and shall see further throughout this volume), computational perspectives are a means of connecting between levels, and it is not necessary to model the whole brain in order to make contributions that can be useful clinically (see, e.g., Kurth-Nelson et al. and Totah et al., this volume).

### How Can the Computational Perspective Handle the Heterogeneity of Real Patients?

Clinical presentations are highly heterogeneous, presenting variability both across patients and within individual patients across time. Totah et al. (this volume) directly raises this issue and discusses it in the relation to individuals. Flagel et al. (this volume) addresses the temporal aspects of this by incorporating trajectories through time directly into the proposed nosology.

Psychiatric disorders tend to have a highly complex temporal trajectory. Although current computational perspectives are capable of integrating temporal trajectories (both recurring and progressing) into their constructs, few current computational models have directly addressed this temporal complexity. How will models be able to capture that temporal trajectory in a way that is informative but not constraining? In a sense, treatment is about changing the trajectory of future symptoms. Can computational models that capture the trajectory of a psychiatric dysfunction be used to guide treatment by making predictions about how those trajectories will change?

Patients have extensive social and psychological lives; the dysfunctions in their computational systems will bleed over into these areas, complicating the picture. It is unlikely that a given patient will have an isolated dysfunction that can be identified as a single failure mode that can be treated by a single paradigm. Is it possible to model the complexity of a given patient? On the other hand, it is an open question whether we need to model the complete patient

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

in order to treat a given dysfunction. Even small improvements in the three challenges (nosology, biomarkers, treatment) and three promises (genetics, circuits, personalized medicine) would be a useful contribution.

**What Is the Appropriate Level of Analysis?**

Another important open question is to ask at what level of analysis the computational perspective should be applied: Do we need to understand the circuit level of a dysfunction? Do we need to understand how genetic variation affects ion channels thus affecting subcircuit interactions leading to changes in computation? Or do we need to understand the computations of whole neural circuits? Humans are fundamentally social animals. Do we need to understand that social computation?

The answer is likely to be that it will depend on the specific psychiatric dysfunction. In general, as noted above, the computational perspective is particularly useful for connecting different levels of understanding. For instance, to understand how a genetic variant can lead to differences in neural responsiveness for an individual neuron type, which can lead to a change in neural circuit dynamics, which can lead to a susceptibility to social or physical stress, one needs to understand the computation being performed at each level. However, if one knows that social interactions depend on decision-making systems, then it may be enough to start from behavioral tests of those decision-making systems.

**What Are the Dysfunctions?**

As noted above, psychiatry has developed categorizations of dysfunction that are reliable but unlikely to be valid. An important open question is whether those categorizations contain enough validity to start from or whether we need to start over with a new nosology. It would obviously be easier to use the current taxonomy of psychiatric dysfunction to bootstrap a (hopefully more valid) computationally and mechanistically justified taxonomy. It is likely that it will be very difficult to completely throw out the current nosology of psychiatric dysfunction (e.g., DSM-IV-TR or DSM-V), but how much it needs to be modified is going to be a critical open question.

An important aspect to this open question is that the answer may be very different when applied to disorders presenting with a limited number of syndromes (e.g., obsessive-compulsive disorder) and when applied to general broad spectrum disorders such as anxiety or depression. It is likely that many disorders with a wide spectrum of behavioral manifestations actually consist of several different disorders that have been categorized together. Schizophrenia, for example, may be an example of a super-category, where many failure modes drive many behavioral outcomes (Silverstein et al. 2013). Additionally, some disorders may be symptoms, whereby many underlying failure modes can lead

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

to the same general outcome (like heart attack and indigestion both causing chest pain). Addiction has been suggested to be such a disorder, where multiple different failure modes can all lead to continued drug use (Redish et al. 2008). Whether computational biomarkers can pull the underlying disorders out or whether there is actually a single underlying dysfunction remains unresolved.

There has been a recent attempt by the U.S. National Institute for Mental Health (NIMH) to create a new categorization of topics for research and analysis called Research Domain Criteria (RDoC) based on psychological constructs, such as attention, cognition, reward systems, etc. (Insel et al. 2010). The RDoC process is still under development, but it is not clear how much of a role computational perspectives have played, or will play, in RDoC's development. The new nosology proposed in this volume offers a novel, nuanced perspective on this difficulty, providing a way of integrating existing taxonomies with dimensional and computational constructs such as those proposed by RDoC.

### What Is the First Exemplar?

At this point, most computational models and explanations for psychiatric dysfunction and treatment have been based on small-scale problems that occur within limited experimental domains, with limited and abstract cue-sets and simple decision components. Although these models have been built at very abstract levels and applied to small-scale (toy) problems, they do capture key factors that have been theorized to drive aspects of psychiatric dysfunction. However, for these models to make an impact on psychiatry, there needs to be a path from computation and theory to clinical practice.

Can we derive a new pathway to understanding psychiatric dysfunction? Is there a general paradigm to apply this computational perspective to psychiatry? A new field, called "computational psychiatry," has begun to emerge (Rangel et al. 2008; Maia and Frank 2011; Montague et al. 2012; van der Meer et al. 2012b; Redish 2013), but it is unclear what that pathway is. The chapters in this book propose first steps toward this new pathway starting from this new computational understanding of psychiatry.

## The Strüngmann Forum

A group of scientists, split evenly between computational neuroscientists and clinical psychiatrists met in June 2015 in Frankfurt under the rubric of the Strüngmann Forum to discuss these issues. The goal of this forum and this book that has arisen from it was to bring together leaders in the fields of psychiatry and computational neuroscience to see if we can make progress on these open questions.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

To take on these questions, we divided into four working groups, each of which addressed a key topic relating to these issues:

1. *Mechanisms*, examining the way that computational perspectives can provide ways of connecting mechanistic differences (genetic variation, differences in social experiences, their interaction) with psychiatric behavioral outcomes
2. *Modeling realistic psychiatry*, examining how computation can address realistic psychiatric patients, who often show comorbidities, who often shift from diagnosis to diagnosis, and who often express complex compensation mechanisms in response to treatments
3. *Nosology*, examining how the computational perspective changes the taxonomy of diagnoses, addressing how the dimensionality of constructs, particularly computational constructs, can be integrated with the clinical practice
4. *A first example*, looking at what it would take to find specific examples to determine whether we have enough at this point to measurably improve treatment for a given dysfunction

The goal of this book is thus to begin to get at the crux of a new question: *How does the computational perspective change psychiatry?* Our hope is that this forum and book can form a concrete framework for future studies and serve as an opening to jump-start this potentially very important cross-field interdisciplinary interaction.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.