Computational phenotyping: using models to understand individual differences in personality, development, and mental illness

Edward H. Patzelt,[1*] Catherine A. Hartley,[2] & Samuel J. Gershman[1]

[1]Department of Psychology and Center for Brain Science, Harvard University

[2]Department of Psychology and Center for Neural Science, New York University

*Corresponding author

In press at Personality Neuroscience

Abstract

This paper reviews progress in the application of computational models to personality, developmental and clinical neuroscience. We first describe the concept of a computational phenotype, a collection of parameters derived from computational models fit to behavioral and neural data. This approach represents individuals as points in a continuous parameter space, complementing traditional trait and symptom measures. One key advantage of this representation is that it is mechanistic: the parameters have interpretations in terms of cognitive processes, which can be translated into quantitative predictions about future behavior and brain activity. We illustrate with several examples how this approach has led to new scientific insights into individual differences, developmental trajectories, and psychopathology. We then survey some of the challenges that lay ahead.

Introduction

The study of personality has a rich history examining individual differences in how we behave, relate to ourselves and each other, and understand our experiences and environment. This work has had the significant challenge of linking multiple levels of analysis spanning complex neural and cognitive processes. Recently, computational models have provided a powerful tool to mathematically formalize this complexity, and provide rich descriptions of the processes underlying human behavior. In the present review, we discuss the concept and promise of a computational phenotype – a collection of mathematically derived parameters that precisely describe individual differences in personality, development, and psychiatric illness.

Traditional approaches to personality are grounded in the study of individuals and how they differ across a range of psychological characteristics that are indexed via measures of traits or symptoms. The most widespread example of this is "general intelligence" (Spearman, 1904). Individuals higher on general intelligence experience better educational (Deary, Strand, Smith, & Fernandes, 2007) and job-related outcomes (Ree, Earles, & Teachout, 1994; Schmidt & Hunter, 2004). However, this research is largely descriptive; general intelligence is a composite measure of several underlying cognitive processes including, but not limited to, working memory (Alloway & Alloway, 2010), verbal and spatial ability, reasoning and processing speed (Deary, Penke, & Johnson, 2010; Hunt, 2011; Lubinski, 2004). This composition of processes has been examined experimentally, but rarely formalized mechanistically. A formal mechanistic definition describes how and why the composition of processes leads to the observable outcome or behavior.

Computational applications to psychiatry have been widely advocated in recent literature (Adams, Huys, & Roiser, 2015; Friston, Stephan, Montague, & Dolan, 2014; Huys, Maia, & Frank, 2016; Huys, Moutoussis, & Williams, 2011; Maia & Frank, 2011; Montague, Dolan, Friston, & Dayan, 2012; Paulus, Huys, & Maia, 2016; Petzschner, Weber, Gard, & Stephan, 2017; Schwartenbeck & Friston, 2016; Stephan, Iglesias, Heinzle, & Diaconescu, 2015;

Stephan & Mathys, 2014; Wang & Krystal, 2014; Wiecki, Poland, & Frank, 2015). In particular, the process of computational phenotyping has been described in considerable depth using real and simulated data (Schwartenbeck & Friston, 2016; Wiecki et al., 2015). Yet, computational perspectives in the fields of personality and development have been relatively limited. Thus, the current review has three goals. First, we broadly outline how computational phenotypes work, and why individuals differ in their phenotype. Second, we review recent work that illustrates the benefits of using computational phenotypes to investigate individual differences. Third, we look forward to challenges in the practical application of computational phenotypes.

Computational Phenotypes: How and Why

A computational phenotype is a set of parameters, derived from neural and behavioral data, which characterizes an individual's cognitive mechanisms. We broadly schematize the process of deriving individual phenotypes in Figure 1A. This more explicit mechanistic characterization complements traditionally descriptive trait and symptom measures in several ways: it formalizes cognitive processes quantitatively, and reduces dimensionality by compressing the target process into a parameter or set of parameters. Moreover, these parameters vary within and between individuals, providing an opportunity to examine individual differences in computational mechanisms. The parameters are also sometimes linked to underlying neurobiological mechanisms.

To illustrate how computational models can provide a mechanistic understanding of behavior, Figure 1B shows a case study of Pavlovian conditioning (Rescorla & Wagner, 1972; Sutton & Barto, 1998). In this experimental paradigm, a light signals the receipt of reward. The Rescorla-Wagner model (Rescorla & Wagner, 1972) explains how the light comes to acquire an association with reward (the "cue value", denoted by V) over the course of conditioning. The learning equation uses the cue value on the current trial t (V_t) to calculate the cue value for the next trial (V_{t+1}):

$$V_{t+1} = V_t + \alpha\delta_t$$

where α is a learning rate parameter (governing how quickly an individual learns) and the reward prediction error δ_t is defined as the cue value from the last trial subtracted from the observed reward r_t.

$$\delta_t = r_t - V_t$$

These equations are formal definitions and serve as mechanistic hypotheses about a wide range of learning and decision-making processes with intricate ties to neurobiology. For example, reward prediction error signals have been found in midbrain dopamine neurons and functional activation in the ventral striatum (Glimcher, 2011; O'Doherty et al., 2004; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006; Schultz, Dayan, & Montague, 1997). In this simple model, the computational phenotype typically corresponds to the learning rate (Figure 1C), which has been linked to genetic (Frank, Moustafa, Haughey, Curran, & Hutchison, 2007) and developmental differences (Christakou et al., 2013; van den Bos, Cohen, Kahnt, & Crone, 2012) between individuals. The prediction error signal itself has sometimes been used as a computational phenotype, distinguishing learners from non-learners (Schönberg, Daw, Joel, & O'Doherty, 2007) and tracking individual differences in the relationship between fluid intelligence and dopamine synthesis (Schlagenhauf et al., 2013). We return to the latter study in depth, within the section on computational phenotyping in personality.

Another key advantage of computational phenotypes, such as learning rate, is dimensionality reduction. Describing a behavioral phenotype without a computational model requires a collection of parameters (e.g., accuracy, reaction time, choice preference) that roughly approximate the process of interest. Computational model parameters compress this

information into a single parameter (e.g., learning rate) or set of parameters that specify how cognitive mechanisms produce behavior and neural activity.

In sum, computational phenotypes define how the cognitive process works mechanistically and provides rich descriptions about why individual variation in phenotypes (e.g., learning rate) produces different behavioral outcomes and neural activity.

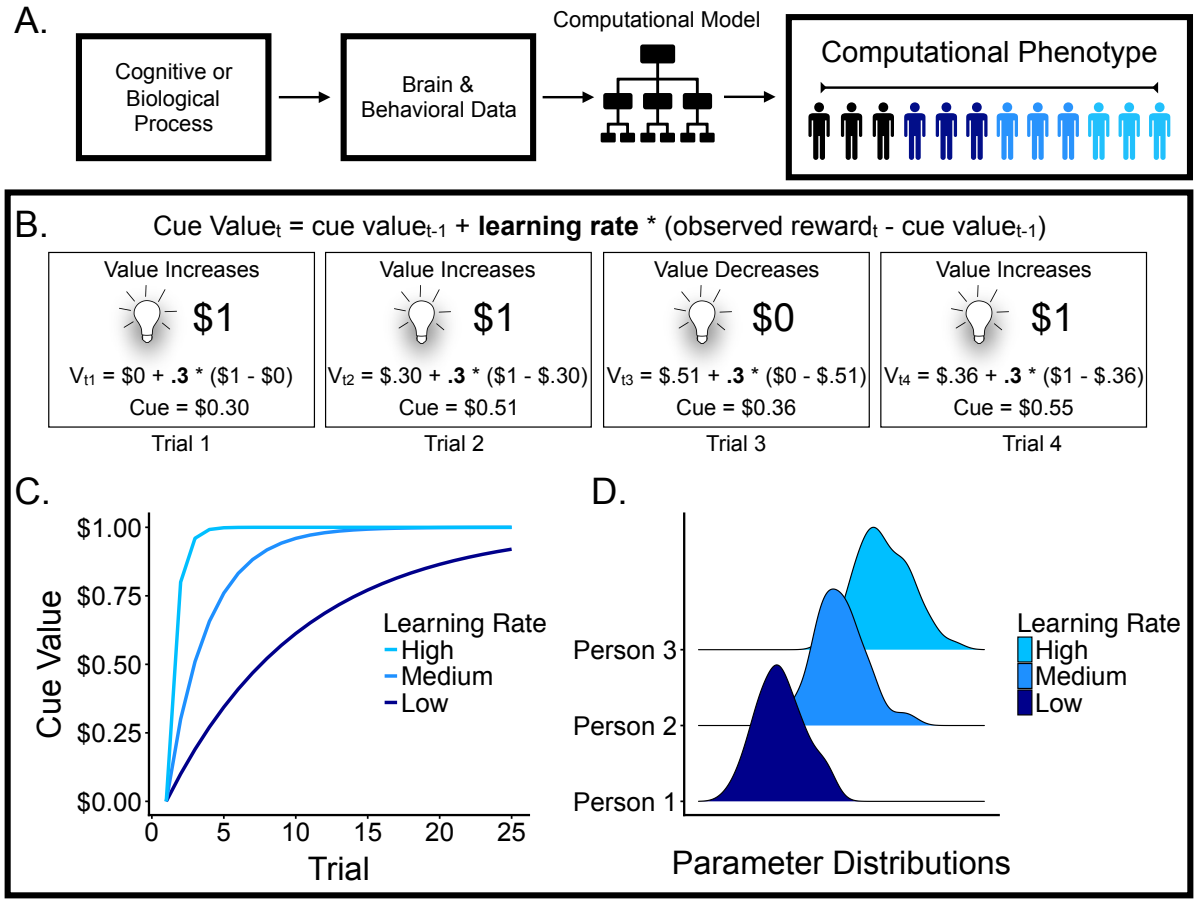

Figure 1: A. Computational phenotyping pipeline. Underlying cognitive or biological processes give rise to brain or behavioral data. The data is entered into the computational model, which produces a set of parameters representing the phenotype. B. Process represented by computational phenotype. In this example, the light represents a cue that indicates a monetary reward. The value of the cue changes on each trial as a function of the value of the cue on the last trial (V_{t-1}), the learning rate (i.e., computational phenotype; 0.3 in the illustration), and the

prediction error (observed reward – cue value$_{t-1}$) (Rescorla & Wagner, 1972). C. Learning rate is the computational phenotype. It varies between individuals, which is why the cue value changes at different rates for each person. D. Learning rates are estimated using Bayesian analysis, increasing parameter sensitivity by using posterior distributions that incorporate uncertainty about the phenotype within and between individuals.

Model Selection and Parameter Estimation

Any study of computational phenotypes faces two methodological questions: how to select the appropriate model, and how to estimate the parameters of that model. Here we will briefly review the main approaches to these questions.

Models are typically evaluated in one of two ways. Goodness-of-fit criteria, such as the likelihood ratio test, the Bayesian information criterion, and the Akaike information criterion, evaluate how well the model fits the data, while penalizing for model complexity. Bayesian model selection criteria are similarly motivated, but place a full distribution over models. Each of these criteria is grounded in different theoretical foundations, so it is often useful to calculate multiple criteria. Predictive criteria evaluate how well a model predicts held-out data. For example, cross-validation uses a model fit to one subset of the data to predict another subset of the data.

Parameter estimation methods fall into one of two categories. Point estimation methods are based on fitting a single set of parameters for each individual. Bayesian methods are based on estimating a posterior distribution over parameters, which allows the researcher to quantify parameter uncertainty (Figure 1D). Hierarchical Bayesian models (see Gelman et al., 2013; Wiecki et al., 2015) take this one step further, estimating distributions over both group-level and individual-level parameter estimates. Researchers can also incorporate prior beliefs about parameter estimates from other datasets, thereby increasing parameter reliability, identifiability, predictive validity, and sensitivity to individual differences (Gershman, 2016).

Computational Phenotyping: Personality, Development, and Psychiatric Illness

Next, we will illustrate the value of computational phenotypes from several different perspectives. Given the scientific breadth of this review, for each perspective we will focus on specific case studies rather than providing exhaustive coverage of the literature. We will show how this approach can reveal new insights into individual differences in personality and examine how the computational phenotype changes over the course of development and aging. Finally, we will show how differences between healthy and disordered brain function can be mapped onto systematic changes in the computational phenotype.

The Computational Structure of Personality

Traditional approaches to the study of personality, such as factor analysis, have been particularly effective in reducing the high-dimensional space of personality to latent constructs such as the Big Five (openness, conscientiousness, extraversion, agreeableness, neuroticism) (Tupes & Christal, 1992). These personality dimensions are largely stable across the lifetime and predict a number of individual differences (e.g., religiosity, dating frequency, and alcohol use among many others; Paunonen, 2003). Despite this predictive validity, traditional personality constructs are largely agnostic as to the cognitive mechanisms by which differences in personality lead to differences in behavior.

For example, conscientiousness is associated with a wide range of adaptive behaviors and outcomes such as greater health and longevity (Bogg & Roberts, 2013), and increased reliability and goal-directed behavior (Jackson et al., 2010). Indeed, how people differ in conscientiousness has been well documented, but it is still relatively unclear as to why people differ in conscientiousness (Abram & DeYoung, 2017). In part, this is due to the fact that conscientiousness is comprised of a heterogenous composition of underlying processes. Disentangling these processes is a task for which computational phenotyping can be uniquely

useful. The specific processes can be operationalized, such as why people higher in conscientious seek more goal-directed behavior. Identifying the computational phenotypes associated with these personality constructs offers the opportunity to link the predictive validity of the construct to its underlying mechanisms. In this section, we examine a set of examples that illustrate what computational models have to offer as a complement to these traditional constructs.

Personality: Goals and Habits

Computational modeling has had an enormous impact on our understanding of decision-making. Here we focus on one particular aspect of this research area: the distinction between two forms of action selection, one based on goals and one based on habits. Initial studies theorized that goal-directed behavior (as studied in rats) was subserved by a "cognitive map" of the environment that supported flexible pursuit of goals (Tolman, 1948). Tolman hypothesized the use of latent learning and planning processes that went far beyond the stimulus-response habits posited by the behaviorists (Thorndike, 1911). Despite the intuitive link to our everyday experience, researchers had only glimpses into the underlying processes. It took more than 50 years to integrate advances in engineering (Bellman, 1957), computer science (Sutton & Barto, 1998), neuroscience (Schultz et al., 1997), and psychology (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Dickinson, 1985) into a synthetic theoretical framework for understanding how the human brain carries out goal-directed and habitual action. This modern computational synthesis conceptualizes goal-directed action arises from using an internal model ("model-based" control) of potential actions and their consequences in the environment, whereas habits arise from a trial-and-error learning system that does not exploit an internal model ("model-free" control).

By constructing explicit computational models of these two systems and their interplay, researchers have been able to capture individual differences in the degree of reliance on model-

based vs. model-free control using a single parameter estimated from a canonical task (Daw et al., 2011). This line of work has led to the study of how stress (Otto, Raio, Chiang, Phelps, & Daw, 2013), age (Decker, Otto, Daw, & Hartley, 2016; Eppinger, Walter, Heekeren, & Li, 2013), and psychiatric illness (Gillan, Kosinski, Whelan, Phelps, & Daw, 2016; Sebold et al., 2014, 2017; Voon et al., 2015) affect, or fail to affect (Nebe et al., 2018), the delicate balance between model-based and model-free control.

Individual variation in model-based control was recently captured by Otto and colleagues when they examined how model-based control is affected by individual differences in stress response (Otto et al., 2013). Participants submerged their arms in ice-cold water (a commonly used acute stress manipulation) and their cortisol levels were measured. Subsequently, they completed a two-step sequential decision task (Daw et al., 2011), that we will refer to as the "two-step task". Computational parameters fit to this task characterize several aspects of learning and decision-making, including the relative contribution of model-free and model-based control for each individual. Otto and colleagues found that participants with higher cortisol levels (greater stress response) exhibited less model-based control. In turn, this effect was modulated by working memory capacity such that greater working memory attenuated stress-induced reductions in model-based control. The key insight from this study is that the precise characterization of how stress and working memory affect individual variation in the computational phenotype (i.e., model-based control), thereby shifting the balance between goal-directed and habitual action. Future work could seek to understand how model-based control does, or does not, covary with conscientiousness and stress.

Personality: Social Cognition

Personality measures such as extraversion and agreeableness are composed of questions about social interaction, including how we relate to ourselves and others. Computational phenotyping increases our understanding of social interaction by specifying the mechanisms

underlying social cognition. For example, computational models of social cognition include parameters representing how quickly we change our view of others, beliefs about the motivations driving their behavior, and a host of other features of social interaction. A recent study (Diaconescu et al., 2014) provides a nice example of computational phenotyping of social cognition in an economic decision-making game.

Diaconescu and colleagues (Diaconescu et al., 2014) used a paradigm where participants were asked to predict the outcome of a lottery. Each participant was paired with an advisor who provided information to aid in the participant's lottery prediction. Importantly, the advisor was incentivized to provide misleading or helpful information, and this varied over time. The critical question for the participant was whether or not to trust the advice of the advisor. Two key parameters from the computational model were (1) a parameter representing the perceived volatility of the advisor's intentions (i.e., how quickly the advice shifted between misleading or helpful), and (2) a parameter representing the perceived advice correctness. When the perceived volatility of the advisor's intentions was high, players weighted their advice lower. Strikingly, players with higher self-reported perspective-taking proficiency had more stable representations of their advisor. This was indicated by slower changes in their belief about advice correctness. Thus, a personality trait (perspective-taking proficiency) directly corresponded to a parameter representing the participant's estimate of another person's trustworthiness. In this example, we have a computational phenotype with parameters for each individual describing *how* and *why* they ultimately decide to take the advice of another person. This computational approach was subsequently extended to the relationship between social cognition and a personality questionnaire measuring autism traits in a healthy population (Sevgi, Diaconescu, Tittgemeyer, & Schilbach, 2016).

Autism is characterized by impairment in social communication and social interaction leading to great difficulty maintaining interpersonal relationships. Moreover, autism traits are continuously distributed in the general population (Robinson et al., 2011).To investigate the

processes that underlie these traits, Sevgi and colleagues employed a computational approach in a social decision-making task while measuring a score on the autism spectrum in a healthy population (Sevgi et al., 2016). They used a game in which using social cue information (indicated by the directional "gaze" of a human avatar) resulted in higher task performance. A computational parameter that represented the weighting of this information in subsequent decisions was correlated with autism score such that higher autism traits were associated with less reliance on social information during decision-making. Moreover, the study showed that individuals high on the autism spectrum showed particular difficulty integrating social advice under more volatile task conditions. Thus, a computational phenotype characterizing a social decision-making process provides a specific mechanism whereby elevations in autism traits are associated with a decreased ability to effectively learn from social information. Next, we turn to the use of computational phenotyping to identify mechanisms underlying individual differences in how people process threatening situations. This is particularly relevant to the construct neuroticism, whereby people higher in this trait experience greater levels of anxiety and worry.

Personality: The Spontaneous Recovery of Fear

A core feature of adaptive behavior is the ability to update our beliefs about threatening situations once they no longer pose a threat. However, some individuals continue to feel fear in apparently safe situations, whereas others seem to learn that a situation no longer poses a threat. In accordance with this idea, a recent paper by Gershman and Hartley (Gershman & Hartley, 2015) demonstrated how a computational phenotype helps explain *why* some people seem to have persistent fears, while others do not.

Gershman and Hartley measured skin conductance response during Pavlovian conditioning. The experiment consisted of three phases: (1) acquisition of the initial fear association by pairing cues with shock, (2) extinction of the fear association by presenting the cues repeatedly without shock, and (3) testing of fear response one day later. Spontaneous recovery of fear was

measured as the difference between skin conductance response on the first block of test relative to the last block of extinction (i.e., how much did an individual's fear response to the cue re-emerge, despite having extinguished this fear response on the previous day). Gershman and Hartley fit a computational model of learning to the acquisition and extinction skin conductance data. This model posited that participants make inferences about the "latent causes" underlying the cue-shock pairs. When the contingencies change sufficiently, the participants should infer that a new latent cause is active. A single parameter controls the sensitivity of latent cause inferences to contingency change. For small values of this parameter, the acquisition and extinction phases are clustered together into a single cause, producing unlearning of the acquired fear and hence no possibility of recovery at test. For large values of this parameter, the acquisition and extinction phases are separated into separate latent causes, thereby protecting the acquired fear from extinction, thus making spontaneous recovery possible.

Using a computational model, Gershman and Hartley clustered participants into two groups on the basis of the sensitivity parameter. As predicted, participants with small sensitivity values apparently unlearned the fear association, showing no evidence of spontaneous recovery. In contrast, participants with larger sensitivity values inferred separate acquisition and extinction latent causes, and accordingly showed spontaneous recovery. Thus, this study demonstrated how a computational phenotyping approach can explain *why* some individuals may continue to feel threatened in environments that no longer pose a threat.

Personality: The Mechanisms of Fluid Intelligence

As noted above, intelligence is comprised of a complex set of underlying processes. A recent study by Schlagenhauf and colleagues (Schlagenhauf et al., 2013) validated complex attention and reasoning as a subprocess of general intelligence using computational modeling. Participants completed a reversal learning task during fMRI and this was followed by a PET scan used to measure dopamine synthesis capacity. It was found that reward prediction errors

in the ventral striatum positively correlated with IQ, and this was specific to the complex attention and reasoning portion of the general intelligence assessment. Moreover, the ventral striatal reward prediction error signal was inversely correlated with dopamine synthesis. Together these findings suggest that a component of the computational phenotype (reward prediction errors) are a promising target for understanding individual differences in fluid intelligence.

The Computational Phenotype Across Development and Aging

Development across the lifespan is associated with profound behavioral and psychological changes. For example, adolescence is characterized by hypersensitivity to social context, vulnerability to emotional arousal, increased impulsivity, and a propensity towards drug and alcohol abuse. Adolescence is also accompanied by neurodevelopmental changes in brain structure (Giedd et al., 1999) and function (Casey, Getz, & Galvan, 2008). The challenge is linking brain and behavior to specific cognitive processes that are tuned differently across developmental stages. Understanding the normative trajectory of these processes can help us to identify atypical developmental trajectories. Moreover, individual differences in these processes arise through a developmental process. Computational phenotypes will allow us to better understand and disentangle the factors that influence individual trajectories.

Development: Model-based Control Across the Lifespan

One particularly important phenotype is the expression of model-based control – the critical ability to evaluate the consequences of our actions. The capacity to prospectively plan actions according to their consequences is starkly contrasted in childhood and adulthood. Requiring significant cognitive resources, model-based control relies on prefrontal structures (Doll, Duncan, Simon, Shohamy, & Daw, 2015; Smittenaar, FitzGerald, Romei, Wright, & Dolan,

2013) known to change dramatically across development (Gogtay et al., 2004). Indeed, there is a shift across development from reliance on impulses to deliberative goal-directed planning (Hartley & Somerville, 2015). This behavioral shift mirrors a neurodevelopmental trajectory whereby prefrontal structures engaged during goal-directed evaluation exhibit a gradual process of integration with subcortical brain structures that can support more automatic behavior (Gogtay et al., 2004). The computational phenotype of model-based control is one way to link changes in brain function and structure to behavioral changes in goal-directed action across development.

Building upon this idea, Decker and colleagues (2016) administered the same two-step task discussed above, using a computational model to estimate the relative balance of model-free and model-based control in a developmental sample. They found a near total absence of model-based control in children ages 8-12. Model-based control emerged during adolescence (ages 13-17) and further strengthened during adulthood (ages 18-25). Extending this work, a subsequent study found that age-related increases in model-based control were mediated by increases in fluid reasoning - the ability to integrate distant concepts to solve problems (Potter, Bryce, & Hartley, 2017). The developmental relevance of these findings is bolstered by evidence that model-based control has been linked to variation in dopamine function (Deserno et al., 2015; Doll, Bath, Daw, & Frank, 2016; Sharp, Foerde, Daw, & Shohamy, 2015; Wunderlich, Smittenaar, & Dolan, 2012) and prefrontal cortex function (Daw et al., 2011; Doll et al., 2015; Smittenaar et al., 2013), both of which are known to change across development (Hartley & Somerville, 2015). Interestingly, while Decker and colleagues found an increase in model-based control from childhood into adulthood, a recent study found that model-based control subsequently decreases in older adults.

Eppinger and colleagues (Eppinger et al., 2013) examined the relationship between model-based control, age, and working memory in a sample of younger adults (mean age: 24) and older adults (mean age 69). Older adults showed less model-based control than younger

adults and this effect was further pronounced by shifting the reward probabilities. They demonstrated that older adults have specific difficulties changing their "cognitive map" of the environment in response to unexpected rewards, whereas younger adults changed their decision strategy and explored the new environment. Like Otto et al. (2013), they found that greater working memory was associated with greater model-based control, but only in younger adults. Moreover, following unexpected rewards younger adults engaged in more strategic exploration of the task structure and older adults tended to perseverate on the previously exploited option. The authors suggest this may be due to a deficit updating expected reward values in older adults. By using a computational phenotype and relating it to other age-dependent processes, the authors demonstrate how phenotypes can be used to examine age-related changes in goal-directed and habitual behavior.

Together, these studies demonstrate how a computational phenotype can be used to trace an arc of cognitive changes across development and through senescence.

Development: Counterfactual Deficits in Adolescence

A core feature of adolescence is difficulty simulating the hypothetical outcomes of decisions. In cognitive science, the consideration of these alternative outcomes is referred to as counterfactual thinking. A recent demonstration of counterfactual deficits in adolescence was accomplished via Bayesian model selection. Palminteri and colleagues administered an instrumental learning task (Palminteri, Kilford, Coricelli, & Blakemore, 2016) and applied three separate computational models. While adolescents were best characterized by a simple reinforcement learning model based upon the Rescorla-Wagner learning rule detailed above, adults were best fit by two more sophisticated models. The first was a counterfactual learning model in which adults incorporated task feedback about unchosen options, and the second was a value contextualization model that allowed adults to learn equally from positive and negative rewards. In contrast to symmetrical reward and punishment learning in adults, adolescents were

less likely to learn from punishment. Therefore, this study identifies three separate computational phenotypes that account for developmental changes in learning and specific process components (e.g., counterfactual learning and punishment sensitivity) that underlie these differences.

Debugging the Brain

Computational modeling provides the advantages in overcoming problems of heterogeneity, comorbidity, and non-specificity in psychiatric nosology (Petzschner et al., 2017; Stephan et al., 2015; Wiecki et al., 2015), providing mechanistic links (i.e. computational phenotypes) between translational neuroscience and applied practice (Friston et al., 2014; Huys et al., 2016; Maia & Frank, 2017; Paulus et al., 2016), and even producing single patient clinical predictions (Stephan et al., 2017). To expand, pathological behavior can be linked to brain disruptions through computational models of distortions in the latent cognitive or biological process. Moreover, specific parameters represent individual components of the process, providing targets for intervention. Computational models also hold promise for linking various types of measurement (e.g. behavior, self-report, brain function) at several levels of analysis. In this section we turn back to model-based control, and examine how this phenotype shows specific relationships with different aspects of psychopathology. We then review work that combines phenotyping with machine learning to aid in the study of schizophrenia, and follow that with an illustration of using Bayesian model comparison to identify two separate neurobiological mechanisms for the phenomenon of synesthesia.

Psychopathology: Model-based Control

A core feature of psychiatric illness is over-reliance on habits at the cost of goal-directed action (Everitt & Robbins, 2005). For example, individuals will often continue a pattern of compulsive drug use despite a stated desire to abstain. The goal of abstinence requires actions

that are commensurate with accurate prospective simulations of the severe consequences of relapse. Due to this phenomenological similarity with the prospective simulation aspect of model-based control (Doll et al., 2015), several studies have investigated the balance between model-free and model-based control in psychiatric illness. Model-based impairment has been found in schizophrenia (Culbreth, Westbrook, Daw, Botvinick, & Barch, 2016), OCD, methamphetamine dependence, and binge eating disorder (Voon et al., 2015).

However, the association between model-based control and problematic alcohol use has been somewhat equivocal and concurrently illuminating. Across the subsequent studies the computational phenotype (balance between model-based and model-free behavior) remains formally consistent, yet the phenotype relates to categorical and trait characteristics of problematic alcohol use differentially. This suggests traditional category-based descriptions of heterogeneous phenomena such as addiction may be further specified with computational phenotypes.

In computational investigations of alcohol use problems some studies have found reduced model-based control in detoxified patients (Sebold et al., 2014) whereas others have not (Sebold et al., 2017; Voon et al., 2015). Despite no reductions of model-based control, Sebold and colleagues (2017) found that reduced medial-prefrontal signatures during model-based decision making predicted relapse in detoxified alcohol-dependent patients. In addition, positive views about the reinforcing effects of alcohol were associated with reduced model-based control in patients who subsequently relapsed (Sebold et al., 2017). Yet, other research has found that model-based control is *not* associated with a range of problematic alcohol use, including binge drinking, onset age for alcohol use, and alcohol consumption (Nebe et al., 2018). Together these studies suggest that model-based impairments may have a more nuanced relationship with alcohol use that traditional methods are not well designed to capture.

Indeed, contemporary views of addiction (Everitt & Robbins, 2005; Kurth-Nelson & Redish, 2012) suggest that individuals will engage in complex reasoning and goal-directed activity to

satisfy a craving. This shifts the view of addiction as simply habitual behavior to a process-based account of drug taking and seeking. Meanwhile, the traditional notion of "addiction as habit" relies on a phenomenological observation that compulsive drug seeking is habitual. This leaves out mechanistic accounts of what drives addictive behavior. Fortunately, a large volume of preclinical and human studies suggest that addiction is comprised of multifactorial disruptions (e.g., cognitive, pharmacological, neural) in the learning process (see this book chapter for theoretical integration of this research; Q. Huys, Beck, Dayan, & Heinz, 2014). Challenges in specifying the mechanisms underlying pathological phenomena can also be partially remedied via dimensional approaches to psychiatric illness.

A large online study by Gillan and colleagues (Gillan et al., 2016) used a transdiagnostic approach to studying model-based control in psychopathology. They applied factor analysis to symptom dimensions comprising mood problems, habitual behaviors, and social functioning. They found that model-based control was reduced in a factor termed 'compulsive behavior and intrusive thought' but was unaffected by anxious depression and slightly improved by social withdrawal. Thus, model-based impairments may be specific to symptoms and traits that cluster together.

While relatively few studies have examined model-based control in psychopathology, computational phenotypes provide a common mathematical foundation for understanding goal-directed deficits. The aforementioned categorical studies ostensibly examined the same process, however they may suffer from nosological problems associated with diagnostic classification and description (Cuthbert & Insel, 2013; Insel et al., 2010). Gillan and colleagues illustrate how we can more accurately conceptualize psychopathological phenomena by shared deficits in a certain process represented by a computational phenotype. Clinicians and researchers alike have observed the transdiagnostic nature of psychopathology, but we have been restricted by lack of formalization of the process and dysfunction within the process. In this regard, computational phenotypes may help shift diagnosis towards a process-oriented

understanding of mental illness whereby deficits in the cognitive process are linked to brain disruptions and behavioral impairments.

Psychopathology: Generative Embedding in Schizophrenia

We have largely focused on mechanistic models that describe how the behavioral or neural data were generated (so-called generative models). These generative models can also be combined with machine learning techniques (e.g., Brodersen et al., 2011). We illustrate this idea with a study that uses machine learning to define psychiatric subgroups in schizophrenia (Brodersen et al., 2014). In contrast to generative models, machine learning approaches are agnostic to mechanism and use the data only to classify subjects as patient or non-patient. However, there is a fundamental problem with this approach. It requires the use of DSM or ICD diagnostic labels. Specifically, the researcher labels training data as patient or control and this is the input for the machine learning algorithm. Using these labels, the machine learning algorithm trains itself to classify the data into patient or control. This approach reifies pre-existing theories about categorical diagnoses. Alternatively, Brodersen and colleagues embed a generative model (rather than labels) of the process giving rise to neural data as the input into a machine learning classifier.

Brodersen and colleagues administered an n-back working memory task to a group of patients with a diagnosis of schizophrenia and healthy controls while they were being scanned with functional MRI. They created a generative causal model of the underlying neuronal dynamics (dynamic causal model; DCM) that gave rise to the fMRI data. The DCM described the network dynamics between the visual cortex, parietal cortex, and dorsolateral prefrontal cortex. Subject-level parameters were derived describing the specific neuronal dynamics for each person and these were subsequently entered into a machine learning algorithm that classified subjects into the schizophrenia or control group. Without any clinical information, the classifier was able to sort the subjects with 78% accuracy. Moreover, within the patient group, the classifier identified three different groups of neural network dynamics. Again, without access

to any diagnostic information, the three classifier groups corresponded to three clinical subgroups as shown by significant differences in negative symptoms. This study is a powerful demonstration of linking a neurocognitive model of working memory to ecologically valid clinical diagnoses via completely data-driven approaches.

Psychopathology: Grapheme-Color Synthesia

Grapheme-color synthesthesia is a perceptual disturbance whereby letters and/or numerals are associated with an experience of color. For example, a person may see a black "4" and perceive the color yellow. Interestingly, there are two broad phenomena that characterize the experience of grapheme-color synthesthetes. Projector synthesthetes perceive the color externally such that the number (e.g., 4) appears in the color yellow. Alternatively, associator synthesthetes experience a strong internal association of the color. In a recent study, van Leeuwen and colleagues (van Leeuwen, den Ouden, & Hagoort, 2011) demonstrate the advantage of using computational phenotypes to identify separable neuronal mechanisms that account for these two types of grapheme-color synthesthesia.

van Leeuwen and colleagues administered a synesthesia-inducing paradigm to a group of known synesthetes during fMRI. They used a dynamic causal model to test two competing hypotheses about visual processing abnormalities that could account for projectors vs associators. They found that neural activity in projectors more closely matched (via Bayesian model comparison) a bottom-up processing stream within the fusiform gyrus whereas associators' neural activity matched a top-down processing stream in the parietal lobe. Therefore, Bayesian model comparison was able to validate that projectors and associators have dissociable computational phenotypes.

Challenges Ahead

Computational modeling is a field that holds promise for grounding individual differences in underlying cognitive and neural mechanisms. However, there are several challenges facing the practical use of computational phenotypes.

One challenge concerns specifying the mechanisms underlying the computational phenotype. For example, model-based control is based upon a number of interrelated cognitive mechanisms including working memory (see Voon, Reiter, Sebold, & Groman, 2017, for a recent review). Therefore, impairments in working memory will correlate with impairments in model-based control (Culbreth et al., 2016) and it will be difficult to tease competing mechanisms apart. One possible solution is administering multiple tasks within the same subject and developing models that capture the overlapping sets of mechanisms across these tasks. These models would derive (for example) parameters that concurrently consider working memory demands and sequential decision making to dissociate the relative contributions of various underlying mechanisms.

A second, related challenge is construct validity. While personality psychology has exerted considerable effort in establishing the validity and robustness of its constructs, computational phenotyping has not yet undertaken such a systematic effort. This is particularly important for several reasons. First, it is well known that parameters in computational models are not always identifiable (Gershman, 2016). This means that parameter values can trade off against one another to produce similar predictions, thereby making it much more difficult to draw conclusions about the mechanism underlying patterns of behavior. This can be partially remedied by parameterizing models in such a way that they do not suffer from identifiability issues, or using data-driven constraints on parameter estimates (Gershman, 2016).

Reliable parameter estimates also require tasks with many trials. This is particularly problematic for patients studies where heterogeneity in the underlying mechanisms and deficits

can produce high variance data. One possible solution is to use hierarchical Bayesian modeling to increase sensitivity to individual differences (Nilsson et al., 2011; Wiecki, Sofer, & Frank, 2013; Wiecki et al., 2015).

Adding to this complexity, it is also unclear how, why, or if, the processes represented by the parameters change over time. This is also important for developmental research, as cross-sectional age differences in a given parameter are assumed to mean that the parameter values within-individual exhibit similar age-related changes over developmental time. However, we do not typically know the test-retest reliability of computational phenotypes because models are rarely fit to multiple datasets from the same subject. This means we have no handle on the contribution of state dynamics to trait measures. This issue can be easily remedied simply by collecting more data; even better, we can measure (or experimentally control) the dynamics of other variables, and thus begin to model state-dependent aspects of computational phenotypes (see Kool, Gershman, & Cushman, 2017, for an example). Test-retest reliability will be especially important for establishing the utility of phenotypes in predicting clinical outcomes and treatment development (Stephan et al., 2017) as we move from translational neuroscience to clinical application (Gold et al., 2012; Paulus et al., 2016).

Another challenge concerns the integration of behavioral and neural data. Computational models are typically fit to behavioral data and then the fitted parameters and latent variables are used in the analysis of neural data. However, recent work has shown how simultaneously modeling neural data (e.g., EEG or fMRI, Cassey, Gaut, Steyvers, & Brown, 2016; Turner et al., 2013; Turner, Rodriguez, Norcia, McClure, & Steyvers, 2016; Turner, Van Maanen, & Forstmann, 2015; Turner, Wang, & Merkle, 2017), or self-report measures (Vandekerckhove, 2014) with behavioral data can lead to greater predictive accuracy and integration of latent cognitive abilities with personality constructs. Other approaches, such as behavioral dynamic causal modeling (bDCM) translate experimental stimuli into neural connections, which in turn, gives rise to behavioral outcomes (Rigoux & Daunizeau, 2015). Effectively, the computational

phenotypes represented by bDCMs are neural networks that operate as neurocomputational mechanisms between environmental inputs and behavioral outputs. Neural models of specific brain regions (e.g., the basal ganglia, Frank, 2005) can also link cellular and systems neuroscience to inform decisions about experimental acquisition of behavioral and brain data. This approach provides biologically plausible mechanisms that account for the neural computations that give rise to behavior (Forstmann & Wagenmakers, 2015). However, application of these models by non-experts poses significant challenges because of the mathematical and programmatic skills required.

Accessible software development is critical for adoption of models by non-computational psychologists and neuroscientists. To date, there are relatively limited software tools available (though, see Wiecki et al., 2013), and those that exist can be difficult for non-experts to use. This problem will be gradually remedied as funding bodies and journals place more stringent requirements on software accessibility. In fact, efforts such as the annual computational psychiatry course already provide open source software for reinforcement learning models, hierarchical Gaussian filters, and drift diffusion models.

Computational models also require mathematical skills that are not easily applied, or understood. Indeed, the application of these models to questions in personality, development, and psychiatry has typically required the integration of skills from multiple researchers with different backgrounds (e.g., personality psychologists and computational neuroscientists). Conferences (e.g., the annual Computational Psychiatry course in London), graduate courses, and potentially graduate degree tracks, could aid in filling these technical and conceptual gaps. In addition, simple steps such as attempts to bridge the language of complementary fields will also be important. For example, the article by Brodersen and colleagues (Brodersen et al., 2014) explicitly describes generative embedding methods in a tutorial aimed at researchers with a clinical background.

Finally, we need more systematic evaluations of the assumptions linking computational phenotypes to behavioral and neural data. Often, researchers run a correlation or regression, looking for simple associations without grappling with the possibility that computational phenotypes could be related to observed data in more complex ways. Clinical psychometricians have extensively studied a range of probabilistic models for understanding how different symptoms and traits are related, ranging from factor analysis to undirected networks (Borsboom et al., 2016; Borsboom, Mellenbergh, & van Heerden, 2004). These same kinds of techniques could be applied to analyzing computational phenotypes. However, the importance of these phenotypes depends upon their predictive validity. This is where longitudinal translational research efforts (Paulus et al., 2016), such as those currently underway in the study of schizophrenia (Gold, 2012; Gold et al., 2012), can validate the ecological and clinical utility of computational models.

Despite these challenges, we are optimistic that computational phenotypes have already begun to bear fruit for personality neuroscience and related fields. We envision a future in which they will be applied to precision medicine approaches (Cuthbert & Insel, 2013; Fernandes et al., 2017; Friston, Redish, & Gordon, 2017), where particular latent processes can be targeted for intervention, and optimized for individual people. Similar interventions could be conceived for the purposes of individualized education and the design of incentive mechanisms for improving financial decision making.

Acknowledgments

This research was supported by NSF CAREER award 1654393 (CAH) and the Harvard Brain Initiative.

References

Abram, S. V., & DeYoung, C. G. (2017). Using personality neuroscience to study personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *8*(1), 2–13. https://doi.org/10.1037/per0000195

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2015). Computational psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery, and Psychiatry*, *87*(1), 53–63. https://doi.org/10.1136/jnnp-2015-310737

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Bellman, R. E. (1957). *Dynamic programming*. Dover, NJ: Princeton University Press.

Bogg, T., & Roberts, B. W. (2013). The case for conscientiousness: Evidence and implications for a personality trait marker of health and longevity. *Annals of Behavioral Medicine*, *45*(3), 278–288. https://doi.org/10.1007/s12160-012-9454-6

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Borsboom, D., Rhemtulla, M., Cramer, A. O. J., van der Maas, H. L. J., Scheffer, M., & Dolan, C. V. (2016). Kinds versus continua: A review of psychometric approaches to uncover the structure of psychiatric constructs. *Psychological Medicine*, *46*(8), 1567–1579. https://doi.org/10.1017/S0033291715001944

Brodersen, K. H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny, W. D., Buhmann, J. M., & Stephan, K. E. (2014). Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage: Clinical*, *4*, 98–111. https://doi.org/10.1016/j.nicl.2013.11.002

Brodersen, K. H., Schofield, T. M., Leff, A. P., Ong, C. S., Lomakina, E. I., Buhmann, J. M., & Stephan, K. E. (2011). Generative embedding for model-based classification of fMRI data. *PLoS Computational Biology*, *7*(6), e1002079. https://doi.org/10.1371/journal.pcbi.1002079

Casey, B. J., Getz, S., & Galvan, A. (2008). The adolescent brain. *Developmental Review*, *28*(1), 62–77. https://doi.org/10.1016/j.dr.2007.08.003

Cassey, P. J., Gaut, G., Steyvers, M., & Brown, S. D. (2016). A generative joint model for spike trains and saccades during perceptual decision-making. *Psychonomic Bulletin & Review*, *23*(6), 1757–1778. https://doi.org/10.3758/s13423-016-1056-z

Christakou, A., Gershman, S. J., Niv, Y., Simmons, A., Brammer, M., & Rubia, K. (2013). Neural and psychological maturation of decision-making in adolescence and young adulthood. *Journal of Cognitive Neuroscience*, *25*(11), 1807–1823. https://doi.org/10.1162/jocn_a_00447

Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M., & Barch, D. M. (2016). Reduced model-based decision-making in schizophrenia. *Journal of Abnormal Psychology*, *125*(6), 777–787. https://doi.org/10.1037/abn0000164

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*(1), 126. https://doi.org/10.1186/1741-7015-11-126

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, *11*(3), 201–211. https://doi.org/10.1038/nrn2793

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological Science*, *27*(6), 848–858. https://doi.org/10.1177/0956797616639301

Deserno, L., Huys, Q. J. M., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A. A., … Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 1595–1600. https://doi.org/10.1073/pnas.1417219112

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., … Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology*, *10*(9), e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *308*(1135), 67–78. https://doi.org/10.1098/rstb.1985.0010

Doll, B. B., Bath, K. G., Daw, N. D., & Frank, M. J. (2016). Variability in dopamine genes dissociates model-based and model-free reinforcement learning. *Journal of Neuroscience*, *36*(4), 1211–1222. https://doi.org/10.1523/JNEUROSCI.1901-15.2016

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, *18*(5), 767–772. https://doi.org/10.1038/nn.3981

Eppinger, B., Walter, M., Heekeren, H. R., & Li, S.-C. (2013). Of goals and habits: Age-related and individual differences in goal-directed decision-making. *Frontiers in Neuroscience*, *7*, 253. https://doi.org/10.3389/fnins.2013.00253

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience*, *8*(11), 1481–1489. https://doi.org/10.1038/nn1579

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of 'precision psychiatry'. *BMC Medicine*, *15*(1).

https://doi.org/10.1186/s12916-017-0849-x

Forstmann, B. U., & Wagenmakers, E. J. (2015). *An introduction to model-based cognitive neuroscience*. New York, NY: Springer. https://doi.org/10.1007/978-1-4939-2236-9

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, *17*(1), 51–72. https://doi.org/10.1162/0898929052880093

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(41), 16311–16316. https://doi.org/10.1073/pnas.0706111104

Friston, K. J., Redish, A. D., & Gordon, J. A. (2017). Computational nosology and precision psychiatry. *Computational Psychiatry*, *1*, 2–23. https://doi.org/10.1162/CPSY_a_00001

Friston, K. J., Stephan, K. E., Montague, P. R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry*. https://doi.org/10.1016/S2215-0366(14)70275-5

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. https://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J., & Hartley, C. A. (2015). Individual differences in learning predict the return of fear. *Learning & Behavior*, *43*(3), 243–250. https://doi.org/10.3758/s13420-015-0176-z

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., … Rapoport, J. L. (1999). Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience*, *2*(10), 861–863. https://doi.org/doi:10.1038/13158

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a

psychiatric symptom dimension related to deficits in goal-directed control. *eLife*, *5*. https://doi.org/10.7554/eLife.11305

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(Suppl 3), 15647–15654. https://doi.org/10.1073/pnas.1014269108

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., … Thompson, P. M. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(21), 8174–8179. https://doi.org/10.1073/pnas.0402680101

Gold, J. (2012). Cognitive neuroscience test reliability and clinical applications for schizophrenia. *Schizophrenia Bulletin*, *38*(1), 103. https://doi.org/10.1093/schbul/sbr173

Gold, J. M., Barch, D. M., Carter, C. S., Dakin, S., Luck, S. J., MacDonald, A. W., III, … Strauss, M. (2012). Clinical, functional, and intertask correlations of measures developed by the cognitive neuroscience test reliability and clinical applications for schizophrenia consortium. *Schizophrenia Bulletin*, *38*(1), 144–152. https://doi.org/10.1093/schbul/sbr142

Hartley, C. A., & Somerville, L. H. (2015). The neuroscience of adolescent decision-making. *Current Opinion in Behavioral Sciences*, *5*, 108–115. https://doi.org/10.1016/j.cobeha.2015.09.004

Hunt, E. (2011). *Human intelligence*. New York, NY: Cambridge University Press.

Huys, Q. J. M., Beck, A., Dayan, P., & Heinz, A. (2014). Neurobiology and computational structure of decision-making in addiction. In A. L. Mishara, P. R. Corlett, P. C. Fletcher, A. Kranjec, & M. A. Schwartz (Eds.), *Phenomenological neuropsychiatry : Bridging the clinic and clinical neuroscience*. New York, NY: Springer.

Huys, Q. J. M., Maia, T. V, & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413.

https://doi.org/10.1038/nn.4238

Huys, Q. J. M., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? *Neural Networks*, *24*(6), 544–551. https://doi.org/10.1016/j.neunet.2011.03.001

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., … Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *Journal of Research in Personality*, *44*(4), 501–511. https://doi.org/10.1016/j.jrp.2010.06.005

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement learning systems. *Psychological Science*, *28*(9), 1321–1333. https://doi.org/10.1177/0956797617708288

Kurth-Nelson, Z., & Redish, A. D. (2012). *Modeling decision-making systems in addiction*. (B. S. Gutkin & S. Ahmed, Eds.), *Computational Neuroscience of Drug Addiction*. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-0751-5_6

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General intelligence,' objectively determined and measured." *Journal of Personality and Social Psychology*, *86*(1), 96–111. https://doi.org/10.1037/0022-3514.86.1.96

Maia, T. V. & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154–162. https://doi.org/10.1038/nn.2723

Maia, T. V. & Frank, M. J. (2017). An integrative perspective on the role of dopamine in schizophrenia. *Biological Psychiatry*, *81*(1), 52–66.

https://doi.org/10.1016/j.biopsych.2016.05.021

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Nebe, S., Kroemer, N. B., Schad, D. J., Bernhardt, N., Sebold, M., Müller, D. K., … Smolka, M. N. (2018). No association of goal-directed and habitual control with alcohol consumption in young adults. *Addiction Biology*, *23*(1), 379–393. https://doi.org/10.1111/adb.12490

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, *55*(1), 84–93. https://doi.org/10.1016/j.jmp.2010.08.006

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–454. https://doi.org/10.1126/science.1094285

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110

Palminteri, S., Kilford, E. J., Coricelli, G., & Blakemore, S.-J. (2016). The computational development of reinforcement learning during adolescence. *PLoS Computational Biology*, *12*(6), e1004953. https://doi.org/10.1371/journal.pcbi.1004953

Paulus, M. P., Huys, Q. J. M., & Maia, T. V. (2016). A roadmap for the development of applied computational psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 386–392. https://doi.org/10.1016/j.bpsc.2016.05.001

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology*, *84*(2), 411–424. https://doi.org/10.1037/0022-3514.84.2.411

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-

dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. https://doi.org/10.1038/nature05051

Petzschner, F. H., Weber, L. A. E., Gard, T., & Stephan, K. E. (2017). Computational psychosomatics and computational psychiatry: Toward a joint framework for differential diagnosis. *Biological Psychiatry*, *82*(6), 421–430. https://doi.org/10.1016/j.biopsych.2017.05.012

Potter, T. C. S., Bryce, N. V, & Hartley, C. A. (2017). Cognitive components underpinning the development of model-based learning. *Developmental Cognitive Neuroscience*, *25*, 272–280. https://doi.org/10.1016/j.dcn.2016.10.005

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *The Journal of Applied Psychology*, *79*(4), 518–524. https://doi.org/10.1037/0021-9010.79.4.518

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

Rigoux, L., & Daunizeau, J. (2015). Dynamic causal modelling of brain-behaviour relationships. *NeuroImage*, *117*, 202–221. https://doi.org/10.1016/j.neuroimage.2015.05.041

Robinson, E. B., Koenen, K. C., McCormick, M. C., Munir, K., Hallett, V., Happé, F., … Ronald, A. (2011). Evidence that autistic traits show the same etiology in the general population and at the quantitative extremes (5%, 2.5%, and 1%). *Archives of General Psychiatry*, *68*(11), 1113–1121. https://doi.org/10.1001/archgenpsychiatry.2011.119

Schlagenhauf, F., Rapp, M. A., Huys, Q. J. M., Beck, A., Wüstenberg, T., Deserno, L., … Heinz, A. (2013). Ventral striatal prediction error signaling is associated with dopamine synthesis capacity and fluid intelligence. *Human Brain Mapping*, *34*(6), 1490–1499. https://doi.org/10.1002/hbm.22000

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*(1), 162–173. https://doi.org/10.1037/0022-3514.86.1.162

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *The Journal of Neuroscience*, *27*(47), 12860–12867. https://doi.org/10.1523/JNEUROSCI.2496-07.2007

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schwartenbeck, P., & Friston, K. J. (2016). Computational phenotyping in psychiatry: A worked example. *ENeuro*, *3*(4), ENEURO.0049-16.2016. https://doi.org/10.1523/ENEURO.0049-16.2016

Sebold, M., Deserno, L., Nebe, S., Schad, D. J., Garbusow, M., Hägele, C., … Huys, Q. J. M. (2014). Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology*, *70*(2), 122–131. https://doi.org/10.1159/000362840

Sebold, M., Nebe, S., Garbusow, M., Guggenmos, M., Schad, D. J., Beck, A., … Heinz, A. (2017). When habits Are dangerous: Alcohol expectancies and habitual decision making predict relapse in alcohol dependence. *Biological Psychiatry*, *82*(11), 847–856. https://doi.org/10.1016/j.biopsych.2017.04.019

Sevgi, M., Diaconescu, A. O., Tittgemeyer, M., & Schilbach, L. (2016). Social Bayes: Using Bayesian modeling to study autistic trait–related differences in social cognition. *Biological Psychiatry*, *80*(2), 112–119. https://doi.org/10.1016/j.biopsych.2015.11.025

Sharp, M. E., Foerde, K., Daw, N. D., & Shohamy, D. (2015). Dopamine selectively remediates 'model-based' reward learning: A computational approach. *Brain*, *139*(2), 355–364. https://doi.org/10.1093/brain/awv347

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption

of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron*, *80*(4), 914–919. https://doi.org/10.1016/j.neuron.2013.08.009

Spearman, C. (1904). “General Intelligence,” objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Stephan, K. E., Iglesias, S., Heinzle, J., & Diaconescu, A. O. (2015). Translational perspectives for computational neuroimaging. *Neuron*, *87*(4), 716–732. https://doi.org/10.1016/j.neuron.2015.07.008

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

Stephan, K. E., Schlagenhauf, F., Huys, Q. J. M., Raman, S., Aponte, E. A., Brodersen, K. H., … Heinz, A. (2017). Computational neuroimaging strategies for single patient predictions. *Neuroimage*, *145*(Pt B), 180–199. https://doi.org/10.1016/j.neuroimage.2016.06.038

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

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Abingdon, UK: Routledge.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Tupes, E. C., & Christal, R. E. (1992). Recurrent personality factors based on trait ratings. *Journal of Personality*, *60*(2), 225–251. https://doi.org/10.1111/j.1467-6494.1992.tb00973.x

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206. https://doi.org/10.1016/j.neuroimage.2013.01.048

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2016). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *NeuroImage*, *128*, 96–115. https://doi.org/10.1016/j.neuroimage.2015.12.030

Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological Review*, *122*(2), 312–336. https://doi.org/10.1037/a0038894

Turner, B. M., Wang, T., & Merkle, E. C. (2017). Factor analysis linking functions for simultaneously modeling neural and behavioral data. *Neuroimage*, *153*(March), 28–48. https://doi.org/10.1016/j.neuroimage.2017.03.044

van den Bos, W., Cohen, M. X., Kahnt, T., & Crone, E. A. (2012). Striatum-medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. *Cerebral Cortex*, *22*(6), 1247–1255. https://doi.org/10.1093/cercor/bhr198

van Leeuwen, T. M., den Ouden, H. E. M., & Hagoort, P. (2011). Effective connectivity determines the nature of subjective experience in grapheme-color synesthesia. *Journal of Neuroscience*, *31*(27), 9879–9884. https://doi.org/10.1523/JNEUROSCI.0569-11.2011

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71. https://doi.org/10.1016/j.jmp.2014.06.004

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., … Bullmore, E. T. (2015). Disorders of compulsivity: A common bias towards learning habits. *Molecular Psychiatry*, *20*(3), 345–352. https://doi.org/10.1038/mp.2014.44

Voon, V., Reiter, A., Sebold, M., & Groman, S. (2017). Model-based control in dimensional psychiatry. *Biological Psychiatry*, *82*(6), 391–400. https://doi.org/10.1016/j.biopsych.2017.04.006

Wang, X.-J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*(3), 638–654. https://doi.org/10.1016/j.neuron.2014.10.018

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

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, *7*(August), 14. https://doi.org/10.3389/fninf.2013.00014

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron*, *75*(3), 418–424. https://doi.org/10.1016/j.neuron.2012.03.042