Compute to learn.

Neural implementation of computations underlying associative learning and decision making

Payam Piray
The research presented in this thesis was carried out at the Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, with financial support from the Netherlands Organization for Scientific Research (NWO).

ISBN
978-94-6284-076-8

Cover
Sajad Beigjani

Design/lay-out
Promotie In Zicht, Arnhem

Print
Ipskamp Drukkers, Enschede

© Payam Piray, 2016

All rights are reserved. No part of this book may be reproduced, distributed, stored in a retrieval system, or transmitted in any form or by any means, without prior written permission of the author.
Compute to learn.

Neural implementation of computations underlying associative learning and decision making

Proefschrift

ter verkrijging van de graad van doctor aan de Radboud Universiteit Nijmegen op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken, volgens besluit van het college van decanen in het openbaar te verdedigen op donderdag 1 december 2016 om 12.30 uur precies

door

Payam Piray
geboren op 7 mei 1985
te Ray, Iran
Promotoren
Prof. dr. Ivan Toni
Prof. dr. Roshan Cools

Manuscriptcommissie
Prof. dr. Markus Ullsperger
Prof. dr. Nathaniel Daw (Princeton University, Verenigde Staten)
Prof. dr. Russell Poldrack (Stanford University, Verenigde Staten)
Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>Dopaminergic modulation of the functional ventrodorsal architecture of the human striatum</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>Impulse control disorders in Parkinson’s disease are associated with dysfunction in stimulus valuation but not action valuation</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>Human choice strategy varies with anatomical projections from ventromedial prefrontal cortex to medial striatum</td>
<td>85</td>
</tr>
<tr>
<td>5</td>
<td>Seeing an angry face impairs learning the value of information in an uncertain world</td>
<td>111</td>
</tr>
<tr>
<td>6</td>
<td>Concluding Remarks</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>153</td>
</tr>
<tr>
<td>Appendix</td>
<td>Nederlandse Samenvatting</td>
<td>169</td>
</tr>
<tr>
<td></td>
<td>Acknowledgement</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>About the author</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>List of publications</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td>Donders Graduate School for Cognitive Neuroscience</td>
<td>180</td>
</tr>
</tbody>
</table>
Introduction
Motivation

In 1950, Alan Turing published a paper to suggest a game, the “imitation game”. This paper provided an objective measure of how the question of “can machines think?” could be addressed (Turing, 1950). This was, in a way, the beginning of artificial intelligence, although the term was invented in 1956, two years after Turing’s untimely and sad death.

Just a few months before that, John Cade, an Australian psychiatrist, serendipitously discovered the sedating effects of lithium carbonate in controlling psychotic excitement (Cade, 1949). This was the beginning of a decade of success for psychiatry in which new medications for several psychiatric disorders were discovered.

65 years later, building on recent developments in artificial intelligence, computational neuroscientists and computer scientists have joined the efforts of psychiatrists to understand mental disorders. Consequently, new questions are arising: can machines simulate mental disorders? If they can, can we discover new treatments for mental disorders by simulating their effects on those machines?

Background

This thesis aims to enhance our understanding of the neural implementation of the computations underlying learning and decision making in the human brain. This effort is part of the emerging field of computational psychiatry (Maia and Frank, 2011; Montague et al., 2012; Friston et al., 2014).

This thesis is grounded in three research domains. First, there are systems neuroscience studies of frontostriatal circuitry indicating that information processing in this circuitry is hierarchically organized within anatomically segregated corticostriatal loops (Alexander et al., 1986; Middleton and Strick, 2002; Draganski et al., 2008a), linked by a spiraling dopaminergic circuitry connecting midbrain and the striatum (Haber et al., 2000; Ikemoto, 2007). Second, this thesis is grounded in cognitive neuroscience studies of Pavlovian/instrumental conditioning and associative learning (Pavlov, 1927; Mackintosh, 1974; Adams, 1982). That literature deals with mechanisms underlying prediction of rewards contingent on some stimuli or actions. Different cortical and striatal regions, within the segregated anatomical corticostriatal loops, have been shown to be implicated in different forms of conditioning (Balleine et al., 2007; Graybiel, 2008). Third, this thesis is grounded in computational neuroscience of reinforcement learning.
learning and Bayesian learning models (Dayan and Abbott, 2001; Chater and Oaksford, 2008), a literature which is closely linked to conditioning by constructing computational models of learning and choice (Barto and Sutton, 1982; Barto, 1995). These models are also linked to neuroscience primarily through seminal works of Schultz and others indicating that midbrain dopamine neurons encode reward prediction errors in terms of reinforcement learning models (Montague et al., 1996; Schultz et al., 1997). Several studies in humans and animals have further strengthened the link between these models and neural activity within the corticostriatal circuitry (McClure et al., 2003; O’Doherty et al., 2004; Daw et al., 2006; Schönberg et al., 2007; Steinberg et al., 2013).

Additional notions derived from systems- and cognitive- neuroscience literatures will be discussed in the next four experimental chapters. In this chapter, I provide an overview of the theoretical notions relevant to this thesis, as derived from computational neuroscience. I wish here to give a general theoretical framework unifying different modeling approaches used in this thesis, and articulate the contributions of this thesis to the emerging field of computational psychiatry. This contribution is theoretically embedded within Bayesian learning theory. This theory provides a powerful and general ground for understanding contemporary attempts in modeling brain functions and its disorders. Note that although computational models used in the next experimental chapters (chapters 2-5) are instances of this general framework, the level of mathematical treatments in each chapter is sufficient for that particular chapter. Here, I wish to clarify the theoretical ground and the implicit assumptions of the experimental chapters.

**On computational psychiatry**

Psychiatry is at a crossroads (Insel et al., 2010; Sahakian et al., 2010; Hyman, 2012; Robbins et al., 2012). Patients are typically diagnosed in late phases of their illness, when diseases are already chronic and relapsing. Diagnosis based on descriptive symptoms is often challenging due to complexity and heterogeneity of the psychiatric symptoms and the high chance of co-morbidity among disorders (McHugh, 2005). Consequently, often there are neither necessary nor sufficient conditions for defining a particular disease (Robbins et al., 2012). Furthermore, most available medications lack proper biological bases and might have severe side effects (Insel et al., 2010). In the last two decade, cognitive and systems neurosciences have advanced our understanding of neurobiological bases of brain disorders using state of the art neuroimaging and preclinical tools. This approach can help psychiatry to go beyond descriptive subjective symptoms reported by patients and
define “objective” cognitive and neural markers of psychiatric disorders (Robbins et al., 2012). Computational psychiatry promises to go further than that, by constructing computational models of brain and behavior to help prevention, diagnosis and treatment of psychiatric disorders, and thus bridging the gap between neural and behavioral markers of brain disorders (Maia and Frank, 2011; Montague et al., 2012; Friston et al., 2014; Stephan et al., 2015a, 2015b).

In this chapter, I identify three distinct levels, each with different goals, critical for being able to computationally model brain disorders of learning and choice. The first level is concerned with constructing computational models of the healthy brain, as any useful theory of brain disorders needs to be based on understanding the healthy brain. The second level is concerned with discovering mappings between endophenotypes and psychological traits. The latter are indexed through parameters (or priors, in a Bayesian sense) of the computational model, thereby “individualizing” the computational model. The basic insight is that individual differences are modeled as differences in parameters modeling individuals. The third level is concerned with computational models of brain disorders. This level provides a computational understanding of how a general functional computational model could go awry when individualized models interact with a particular context.

1. Computational modeling of learning and choice

Computational neuroscience has successfully provided formal theories to explain behavior and brain functions. There are several approaches, at different levels of explanation, and with different assumptions. One influential approach is concerned with system-level computational explanation of brain and behavior assuming that there are computable functions (in computer science sense (Turing, 1937)) between brain states and behavior (Montague et al., 2012). This approach is the core of computational psychiatry. There are two comprehensive computational frameworks of learning and choice: Bayesian decision theory encompassing reinforcement learning (Körding, 2007; Dayan and Daw, 2008) and active inference framework (Friston and Stephan, 2007; Friston, 2010). Here, I give an overview of each theory.

Bayesian decision theory

It is an intuitive idea that animals make choices to maximize their utility. This is the basis of Bayesian decision theory, which specifies that animals should make choices to maximize their expected utility, averaging over all their uncertainties (Dayan and Daw, 2008). The term “uncertainty” plays a critical role in this formulation, as real-world situations are rife with different types of uncertainties. In a simple case, consider the reinforcement learning problem of learning the value of different states in the environment, where each state of the world
corresponds to some (probabilistic) reward. The animal makes choices and each choice changes the state of the world and results in another reward. The value of the current state could be defined as cumulative expected future reward discounted over time. However, in the real-world problem, there are uncertainties in reward and in transitions between states. Reinforcement learning theory provides real-time approximations of the subjective value of each state-action pair, \( V(s,a) \), given learning parameters such as learning rate specifying the speed of learning value by each reward experience and discount factor specifying the cost of delay in collecting reward (Sutton and Barto, 1998).

In the real-world situations, however, it is often the case that the state of the world is uncertain. More specifically, there are perceptual uncertainties and animals should make choices taking into account these uncertainties (Körding, 2007; Dayan and Daw, 2008). In other words, animals have only access to their own actions and sensory observations, which are caused by the state of the world (Dayan et al., 1995). In a probabilistic setting, this means that the state of the world is a latent variable and an animal has an internal probabilistic model of its own sensory observations given the world (latent) state and its action, \( p(o | s,a) \), where \( o \) denotes sensory observations. To make choices, the animal needs to solve an inverse problem, which is inferring the state of the world according to sensory observations, \( p(s | o,a) \). This could be computed using the Bayes rule:

\[
p(s | o,a) = \frac{p(o | s,a) p(s | a)}{p(o | a)} \quad \text{Equation 1}
\]

where \( p(s | a) \) is the prior belief, before making sensory observations, about the state of the world given the action. Computing the normalizing term, \( p(o | a) \), is possible by marginalizing over latent state:

\[
p(o | a) = \int p(o | s,a) p(s | a) ds \quad \text{Equation 2}
\]

where \( p(o | a) \) is called marginal distribution and could be conceived as a measure of how well the animal could predict its own sensory observations given its action. This is because the integral in this equation is indeed a weighted average of the predicted probability of the animal’s sensory observations \( p(o | s,a) \) over all possible states (causes) weighted by the a priori belief that that state be the true state of the world. The value-based decision system should take into account the uncertainty in animal’s knowledge about the state of the world by computing an expected value of action:
\[ \hat{V}(a) = \int V(s,a) p(o,a) ds \]  

Equation 3

where \( \hat{V}(a) \) is the expected value of action under the perceptual belief about the state of the world. The integral in this equation could be seen as a way to weight a value computed using a particular world state with the belief that that state is the true state given sensory observations.

Equations 1 and 2 construct perception, i.e. inferring the state of the world given observations, and equation 3 is the basis for choice. In practice, however, there is a critical issue as both integrals in Equations 2 and 3 are often intractable. Similar issue arises for dealing with other forms of uncertainty, for example uncertainties in motor implementation of actions. This requires some forms of approximation. A class of approximations, called variational inference, turns out to be useful for solving these issues (Hinton and Zemel, 1994; MacKay, 1995; Jordan et al., 1999). The basic idea is to approximate the posterior distribution with a parametric “variational” distribution, such as Gaussian, and optimize the parameters of the variational distribution to make it more similar to the posterior. In addition, this method provides a good lower bound of the marginal distribution, \( p(o) \). If \( q(s) \) is the variational distribution over the latent state of the world, it can be shown that:

\[ \log p(o \mid a) = -F(q(s),a) + D(q(s), p(s \mid o,a)) \]  

Equation 4

where \( F(q(s),a) \) is called variational bound or free energy that depends on the variational distribution and actions and \( D(q(s), p(s \mid o,a)) \) is called Kullback-Leibler divergence that could be understood as a measure of dissimilarity between \( q(s) \) and \( p(s \mid o,a) \). The mathematical definitions of these quantities are not important here (see Supplementary equations). However, there are three key features about this equation and these quantities, which should be highlighted. First, whereas calculating the free energy requires to have the internal model, \( p(o \mid s,a) \), calculating \( D(q(s), p(s \mid o,a)) \) needs the posterior, \( p(s \mid o,a) \) (see Supplementary equations). Therefore, working with free energy is possible without solving the integral in Equation 2. Second, since divergence (like other dissimilarity measures) is always positive, negative free energy provides an approximation (more precisely a lower bound) of \( \log p(o \mid a) \). Indeed these two quantities are equal if \( D(q(s), p(s \mid o,a)) = 0 \), which only happens if \( q(s) \) is equal to the posterior probability of the state of the world, \( p(s \mid o,a) \). Third, while both the negative free energy and divergence are functions of the variational distribution, \( q(s) \), their sum is...
independent of the variational distribution. Therefore, any change in the negative free energy with respect to the variational distribution necessarily reflects a similar change in the divergence in opposite direction as their sum is independent of changes in the variational distribution.

From these three features, we conclude that the negative free energy can be maximized with respect to parameters of the variational distribution, \( q(s) \), which necessarily reduces the divergence (as their sum is independent of \( q(s) \)) and provides an approximation for the posterior, \( p(s \mid o,a) \). Therefore, by choosing “appropriate” variational distributions, the animal can solve the perception problem (inferring environmental state based on sensory observations).

This variational approximation also provides a promising approach to efficiently approximate the intractable integral in Equation 3, arising in computing expected value of state-action pairs. Specifically, by replacing \( p(s \mid o,a) \) with \( q(s) \) in Equation 3, one could compute an approximate expected value. The limitations is that the variational distribution, \( q(s) \), should be chosen in a way to also simplify computing the integral in Equation 3 given the form of value function, \( V(s,a) \).

**Active inference framework**

The active inference framework extends the variational treatment of perception to the domain of action selection (Friston and Stephan, 2007; Friston, 2013). To explain this framework, we should see the whole process of perception and action in an iterative manner: animals infer the state of the world based on their sensory input (perception) and sample the world (i.e. change their sensory observations) given the inferred state of the world (action). Therefore, every action is followed by a perception process because the sensory observations have been changed. For example, eye movements sample the environment and thereby change the visual inputs, which then could change our perception of the world. From this perspective, action is intimately linked with perception.

Now suppose that the animal has approximated the current state of the world by evaluating the variational distribution, \( q(s) \). We already mentioned that \( -F(q(s),a) \) is an approximation of (log-) sensory evidence, \( p(o \mid a) \), which is the animal’s ability to predict its own sensory inputs given the action. If we assume that a goal of the animal is to be able to better predict its own sensory observations, then one way to change the sensory input is to make actions that increase \( p(o \mid a) \). Since free energy is a tractable approximation of this quantity, animal could minimize the free energy with respect to action, \( a \), which necessarily maximizes predictability of sensory observations, \( p(o \mid a) \).
We have conceptualized the free energy as an approximation strategy that animals or brains could use to simplify computations related to perception and action. Friston goes beyond this conceptualization and proposes that free energy minimization is the underlying principle of brain computations. He has put forward the hypothesis that computations in the brain minimize time average of free energy (Friston, 2013). This unifying proposal is based on the intuition that biological systems are evolved to live within a bound and any surprising sensory state, those states that are \textit{a priori} unlikely, should be avoided (Friston and Stephan, 2007). Based on this principle, Friston rejects the notion of extrinsic value and suggest that expected value of action (cf. Equation 3) is only the negative expected surprise of sensory observations under approximate perceptual distribution (Friston et al., 2009).

Bayesian decision theory and active inference theory are often treated as inconsistent viewpoints. We showed that these concepts are deeply related, although we admit that there are important conflicting aspects. The relation comes from the practical fact that both theories, like any other Bayesian framework, take into account uncertainties and, consequently, should deal with some generally intractable integrals, which makes approximations necessary. From this perspective, both theories view the brain as an approximate Bayes-optimal. These approximations usually map a difficult inference problem (i.e. computing intractable expectations) to a supposedly easier optimization problem typically with respect to dynamic variables of some tractable distributions such as Gaussian. The dynamic variables, such as the mean and variance of Gaussian, are dependent on some parameters. Here, the difference between dynamic variables and parameters is in the temporal scale of their change. Dynamic variables tend to change rapidly, for example during one task, while parameters tend to change slowly, e.g. during development. The parameters might be defined as constant parameters, or as some prior probability distributions in the Bayesian sense. Since constant parameters are a special case of probability distribution, I refer to parameters and/or their prior generally as priors. In the next section, I discuss the role of these priors in brain computations.

2. Computing vulnerability: mapping traits to priors
Most psychiatric disorders show high heritability (Meyer-Lindenberg and Weinberger, 2006). Yet, we know very little about the mechanisms through which risks for psychiatric disorders might be inherited. In psychiatry, the concept of endophenotypes has been introduced as a mediating factor between predisposing genes and clinical symptoms (Gottesman and Shields, 1973; Gottesman and Gould, 2003). Recently, the concept of cognitive endophenotypes has been suggested as
quantitative and objective neuro-cognitive traits predisposing brain disorders (Robbins et al., 2012). I believe that an important challenge for computational psychiatry is to formally define cognitive endophenotypes in terms of "computational traits", using (generative) computational models of brain and behavior. It has been generally assumed that such a mapping is possible and initial attempts for finding the map, especially in patients (Brodersen et al., 2014), have already been started. Here, I wish to give a theoretical flavor of assumptions behind this approach, as mapping cognitive and neuroanatomical traits into elements of computational models is central to this thesis.

One fundamental powerful feature of Bayesian systems is that they could accumulate knowledge in the form of priors. Regarding the brain, some parts of this knowledge are accumulated through interaction with the world, either through natural selection across generations, or by individual learning. A Bayesian perspective suggests that priors are optimized in relation to environmental possible states. In other words, we can regard priors in the brain as "good" priors given the environment generating data.

It is intuitive that good priors are both informative and sufficiently flexible, enabling new data to change the posterior. This corresponds to the concept of Bayesian model evidence, which has a built-in tradeoffs between fit to the data (e.g. sensory observation, rewarding outcomes etc.) and model complexity (MacKay, 2003). We can obtain some insights into the concept of model evidence in a simple situation, when prior distribution is a uniform within a bound, so that: \( p(w) = 1/w_0 \). Then, it is easy to show that for a Gaussian posterior distribution with uncertainty, \( w \), (log-model) evidence is:

\[
\text{Evidence} = \text{fit} - \text{complexity}
\]

where complexity = \( \log(w_0 / w) \) (MacKay, 2003). Note that complexity is always positive because \( w_0 > w \), as one cannot be less certain about a random variable after observing some samples of. Also it could be easily seen that the complexity term is higher for models with larger a priori uncertainty, \( w_0 \). It is because a higher prior uncertainty corresponds to wider prior distribution, which is capable of making lots of predictions but necessarily with lower probability (Figure 1).

This simple situation can give us some hints about how priors might corrupt inference. Given a fixed internal model, inference can be corrupted due to "bad" priors in two ways: 1) The prior is too wide resulting in a too complex model that is able to make a wide variety of predictions. Such a model is probably too slow and
Introduction

1) The prior is too wide resulting in a simple, fast and easy-to-be-adjusted model. Such a model is probably too rigid and fails to learn a wide range of data. Speculative examples of bad priors are given by Parkinson’s disease and by behavioral addictions. Whereas behavioral addictions are associated with excessive habitual control of action (Everitt et al., 2008), Parkinson’s disease might force patients towards a progressive and excessive reliance on goal-directed control of actions (Redgrave et al., 2010). It seems that both disorders correspond to the same prior, probably in the domain of action and motor control. Behavioral addictions might have excessively narrow priors, resulting in maladaptive rigidity of actions. Parkinson’s disease might have excessively wide priors, resulting in slow and effortful actions. Interestingly, physiologically and psychologically, these brain disorders show opposite properties. While Parkinson’s disease is the pathological state of loss of striatal dopamine, addiction is generally associated with excessive striatal dopamine. Psychologically, it has been reported that non-medicated Parkinson’s disease patients typically do not engage in impulsive or addictive behaviors and tend to not smoke cigarettes or drink alcohol (Dagher and Robbins, 2009).

**Figure 1** A) Three prior distributions over one parameter, w, corresponding to three models with the same internal structure and different priors. Larger prior width corresponds to a more complex model. B) In addition to the priors in panel A, the likelihood function for two observed datasets, D1 and D2, is depicted. Model M1, associated with the red prior, predicts D1 well but fails to predict D2. Model M3, associated with the blue prior, predicts both D1 and D2 equally well but with low evidence, as this model a priori predicts that a wide range of parameters, w, might be responsible for generation of data. Model M2 is the best model, which fits to both dataset well and has lower complexity than M3.
Therefore, if priors are optimized during evolution according to probable world states, then it makes sense to regard heritable traits as priors. This view essentially models individual differences as differences in priors modeling individuals. This perspective also highlights the role of “environment” in development of psychiatric disorders, as trait priors implicitly reflect (evolutionarily-learned) expectations about the world, and they become suboptimal only in contexts violating those expectations. Note that this perspective assumes that two other components of approximate-Bayesian inference, the internal (generative) model and approximations, are the same across individuals. Although untested, this is a general assumption shared by all available computational frameworks of brain and behavior.

3. Modeling disorders of brain
An approximate-Bayesian system relies on three major components, namely priors, generative model of the world and approximations. It has been discussed extensively how these components can model maladaptive behaviors in the context of Bayesian decision-theoretic systems (Huys et al., 2015). Here, I focus on two ways to model pathological computations in a Bayesian model of brain.

The first way is to assume that implementation of the generative model or its approximations is corrupted in some situations. A good example of this view is Redish (2004) model of drug addiction. In this model, drugs are assumed to hijack the reward system by adding a non-compensable drug-induced term to reward prediction errors presumably signaled by dopamine after drug taking (Redish, 2004). Therefore, the value of drug taking action grows and ultimately cancels out any potential cost associated with drug taking. Although such models could be computationally and mechanistically useful, they do not link endophenotypes with possible pathological computational processes and have limited diagnostic value.

Another possibility through which a computational system can go awry is that some individuals with specific priors act (or perceive) maladaptively in some environmental contexts. Although this is unlikely to result in a chronic pathological condition, a severe problem might arise when the consequences of those maladaptive actions exacerbate the same priors leading to them. In the language of control theory, this is a positive feedback loop. Therefore, this approach is grounded in potential endophenotypes associated with a specific mental disorder. We exploit findings from the literature of drug addiction and obesity to give an example. Preclinical animal models of cocaine addiction have shown that animals with low density of D2 dopamine receptors in the ventral striatum before drug exposure are highly impulsive (Dalley et al., 2007), escalate their cocaine
intake after drug exposure (Dalley et al., 2007) and are more vulnerable to develop compulsive addiction-like cocaine seeking (Belin et al., 2008). Furthermore, knockdown of striatal D2 receptors in rats induced compulsive eating in obese rats (Johnson and Kenny, 2010). Additionally, it has been repeatedly observed that drug abuse and obesity progressively reduce striatal D2 receptors (Porrino et al., 2004a, 2004b; Volkow et al., 2008). Therefore, ventral striatal D2 receptors provide a basis for modeling individual vulnerability to addiction.

Building on these findings, we have previously proposed a reinforcement learning model of addiction, where addiction arises due to an imbalance between a ventral striatal system computing prediction errors and a dorsal striatal system implementing action valuation and selection. Specifically, we have assumed that there is a biological bound (prior) on the ability of D2 receptors in one region to tolerate dopaminergic stimulation and overstimulation of these receptors further reduces their density. Thus, those individuals with lower amount of D2 receptors have lower bound and are vulnerable to excessive dopamine release by highly palatable drugs or foods. By assuming that dopamine receptors modulate learning rate based on observations in humans (Frank et al., 2004, 2007b) and monkeys (Groman et al., 2011; but see Piray, 2011), we showed that an imbalance arises between updating of stimulus values responsible for learning and action values responsible for action selection.

This example provides some insights into how a biological system modeling approximate-Bayesian system might go awry. Indeed, interaction between appropriate contexts and bad priors might violate assumptions behind the approximations. In this example, the prior assumptions about maximum amount of prediction error are systematically violated by highly rewarding substances and the violation itself increased the chance of occurrences of more violations. Similar problems could arise in the context of variational approximation.

**Conclusion**

In this work, I have highlighted three distinct levels within the computational psychiatry program emphasizing how Bayesian approaches enable us to model cognitive and neural markers predisposing organisms for brain disorders. Specifically, mapping cognitive and neuronal traits into parameters and priors of learning models enable us to construct “computational traits” with potential diagnostic value. Furthermore, combining priors with internal model enables us to investigate the effects of parameters on learning and choice, which is necessary for understanding pathological computational processes.
Outline of the thesis

In this thesis, I employ computational models, structural and functional neuroimaging tools, and pharmacological interventions, in both healthy and altered human brains, to investigate fronto-striatal circuits. The goal is to understand the computations implemented by different elements of those circuits. This thesis is a step within a long-term research program aimed at defining the cognitive and neural mechanisms that might predispose individuals towards impulsive and compulsive disorders, i.e. disorders known to arise from alterations in fronto-striatal circuits. Dysfunction of elements of those circuits is also associated with disorders such as depression, Parkinson's disease, and schizophrenia.

The primary neural alteration in drug/behavioral addictions is within the mesostriatal dopaminergic system. Importantly, only a small subset of people tend to develop compulsive behaviors, even among those repeatedly exposed to drugs (O’Brien et al., 1986; Deroche-Gamonet et al., 2004; Everitt et al., 2008). Evidence from work with behaving rodents indicates that rats with lower dopamine D2 density in the ventral striatum are impulsive and vulnerable to develop compulsive drug seeking behaviors (Dalley et al., 2007; Belin et al., 2008), which depends on dopaminergic spiraling connections between the striatum and the midbrain (Haber et al., 2000; Belin and Everitt, 2008). Trait impulsivity is hypothesized as a cognitive endophenotype predisposing compulsive drug seeking (Robbins et al., 2012). In chapter 2, we focus on the architecture of the human striatum, its D2-dependent dopaminergic modulations and its relation to trait impulsivity in healthy human participants.

Parkinson's disease is a degenerative neural disorder primarily affecting the motor system. Although Parkinson's disease is physiologically and psychologically opposite to behavioral addictions (Dagher and Robbins, 2009), a subpopulation of patients may become addicted to their own medications or exhibit impulse control disorders, such as pathological gambling, binge eating, excessive shopping and hypersexuality (Voon et al., 2007). In chapter 3, we study maladaptive learning processes in Parkinson's disease patients with impulse control disorders.

Modern systems- and cognitive- neuroscience theories of addiction indicate that addiction arises as a progressive transition from goal-directed voluntary drug-seeking to maladaptive habitual drug-taking behavior (Everitt et al., 2008). In terms of neural control, there is a corresponding transition from the affective and cognitive loops of corticostriatal circuitry to the sensorimotor corticostriatal loop. This transition is mediated by dopaminergic spiraling connections between the
striatum and midbrain (Everitt et al., 2008). Goal-directed and habitual modes of action selection have been hypothesized to reflect model-based and model-free reinforcement-learning computations in the brain, respectively (Daw et al., 2005, 2011). In chapter 4, we investigate the components of corticostriatal circuitry predicting individual differences in exerting model-based and model-free control on actions.

Maladaptive emotion processing and emotion regulation are strongly linked to drug and behavioral addictions especially in initiation of drug use and in relapse (Khantzian, 1997). Emotions are hypothesized to modulate learning and choice, by modulating the saliency of cues predicting reward or punishment (Pearce and Hall, 1980; Phelps et al., 2014). Bayesian and reinforcement learning theories predict that emotions might modulate learning rate (Li et al., 2011b), which changes beliefs about action-outcome contingencies in volatile environments (Behrens et al., 2007b). In chapter 5, we investigate the behavioral and neural mechanisms by which emotions modulate associative learning in humans.

In chapter 6, I summarize findings and discuss future works.
Supplementary equations

The free energy and divergence quantities in Equation 4 are:

\[ F(q(s), a) = - \int q(s) \log \frac{p(s | o, a)}{q(s)} \, ds \]

\[ D(q(s), p(s | o, a)) = - \int q(s) \log \frac{p(s | o, a)}{q(s)} \, ds \]

It could be seen that whereas the posterior, \( p(s | o, a) \), is needed for calculating the divergence, \( D \), the free energy, \( F \), could be computed based on the prior, \( p(s | a) \), and the internal model, \( p(o | s, a) \).
Dopaminergic modulation of the functional ventrodorsal architecture of the human striatum

This chapter is adapted from:
Abstract

Interactions between motivational, cognitive and motor regions of the striatum are crucial for implementing behavioural control. Work with experimental animals indicates that such interactions are sensitive to modulation by dopamine. Using systematic pharmacological manipulation of dopamine D2 receptors and resting-state functional imaging, we defined the functional architecture of the human striatum and quantified the effects of dopaminergic drugs on intrinsic effective connectivity between striatal subregions. We found that dopamine modulates interactions between motivational and cognitive regions, as well cognitive and motor regions of the striatum. Stimulation and blockade of the dopamine D2 receptor had opposite (increasing and decreasing) effects on the efficacy of those interactions. Furthermore, trait impulsivity was specifically associated with dopaminergic modulation of ventral-to-dorsal striatal connectivity. Individuals with high trait impulsivity exhibited greater drug-induced increases (after stimulation) and decreases (after blockade) of ventral-to-dorsal striatal connectivity than individuals with low trait impulsivity. These observations establish a key link between dopamine, intrinsic effective connectivity between striatal subregions, and trait impulsivity.
**Introduction**

The striatum subserves many functions, ranging from incentive motivation to goal-directed action selection and habitual response control, and is implicated in a wide range of neuropsychiatric disorders such as addiction, attention deficit hyperactivity disorder and Parkinson's disease. These various striatal functions have long been thought to depend on information processing within relatively segregated motivational, cognitive and motor regions of the striatum (Alexander et al., 1986). However, recent evidence has highlighted an important functional role for interactions between these different regions of the striatum (Haber and Knutson, 2010; Aarts et al., 2011). For example, according to current theories of addiction, the transition from impulsive to compulsive drug use corresponds to a transition from ventral to dorsal striatal control of drug seeking behaviour (Everitt et al., 2008). Despite the importance of these hierarchical intra-striatal interactions, they have received little attention in human research.

Neuroanatomical data from non-human primates have suggested that the communication between striatal regions is subserved by a network of spiraling connections between the dopaminergic midbrain and the striatum (Haber et al., 2000). Thus dopamine is ideally suited for mediating information flow along the mediolateral striatal gradient through serial reciprocal connections between the striatum and the midbrain. This hypothesis concurs with evidence from work with behaving rodents indicating that the transition from impulsive to compulsive drug use, and the corresponding transition of behavioural control from the ventral to the dorsolateral striatum, can be promoted by dopamine (Dalley et al., 2007; Belin and Everitt, 2008; Belin et al., 2008).

Here we aim to assess how dopamine modulates intrinsic human striatal connectivity by administering dopaminergic drugs and by exploiting inter-individual variability in the direction and extent of dopaminergic drug effects (Cools and D’Esposito, 2011). Individual differences in the personality trait of impulsivity have been shown to correspond with individual differences in dopamine receptor availability (Dalley et al., 2007; Buckholtz et al., 2010) and with the effects of dopaminergic drugs on striatal function (Cools et al., 2007b). Specifically, trait impulsivity is associated with low D2-receptor density in the ventral striatum (Dalley et al., 2007). Moreover, trait impulsivity has been shown to promote the transition of control of reward-seeking behaviour from ventral to dorsal striatal regions (Belin and Everitt, 2008; Belin et al., 2008). Based on this literature, we predict that administration of D2-receptor drugs to healthy volunteers will alter the influence of the ventral striatum on the dorsal putamen, through the dorsal
caudate nucleus, in a manner that depends on trait impulsivity. Following prior work (Cools et al., 2007b; Dalley et al., 2007; Clatworthy et al., 2009), we hypothesize that dopaminergic drugs would have greater effects in high-impulsive subjects (with putatively low D2-receptor density) than in low-impulsive subjects.

We employed a 2x2 factorial pharmacological design and manipulated dopamine receptor stimulation in a group of healthy participants by administration of a dopamine D2-receptor agonist (bromocriptine), a dopamine D2-receptor antagonist (sulpiride), a combination of the agonist and antagonist, and a placebo, in a four-session, within-subject, double-dummy, placebo-controlled cross-over design. This factorial pharmacological design allowed us to assess the neurochemical specificity of effects. If any effects of the D2-receptor agonist bromocriptine depend on dopamine’s action on D2-receptors, then they should be blocked by pretreatment with sulpiride.

Blood oxygenation level dependent (BOLD) signal was measured using functional magnetic resonance imaging (fMRI) during rest. This approach allowed us to relate task-independent features of intrinsic striatal connectivity to trait-related individual differences in mesostriatal dopamine systems. We used stochastic dynamic causal modeling (DCM) (Friston et al., 2011; Li et al., 2011a; Daunizeau et al., 2012) to model interactions between motivational, cognitive and motor portions of the striatum and their modulation by dopamine. This method estimates the extent to which fluctuations in activity of one region cause fluctuations in another region. The results demonstrate that dopamine modulates intra-striatal connectivity and that the degree of dopaminergic modulation of dorsomedial striatum (dorsal caudate nucleus) by ventral striatum is associated with trait impulsivity.

Methods

Participants
Twenty-eight participants gave informed consent approved by the local ethical committee (“Commissie Mensgebonden Onderzoek”, Arnhem-Nijmegen, number: 2008/078). Three participants were excluded from the analysis: two participants withdrew before completing all four sessions; one dataset was unusable due to a technical problem. The 25 participants included in the analysis were right-handed (13 women; mean age 22 years, range 18-30 years), with no relevant medical or psychiatric condition 3 years prior to testing.
Factorial pharmacological design
We employed a 2x2 factorial pharmacological design (Table 1). The two pharmacological factors were bromocriptine, a dopamine receptor agonist, and sulpiride, a dopamine receptor antagonist. Each of these factors could be ‘ON’ (drug) or ‘OFF’ (placebo). Each participant was tested on each cell of this factorial design, receiving 2 different opaque gelatin capsules on 2 separate time points on each of the four testing sessions, corresponding to these two pharmacological factors (double-dummy design). Thus, two pharmacological factors could affect mesostriatal system: whether sulpiride was ON or OFF, and whether bromocriptine was ON or OFF. This design allowed us to quantify not only the main effects of sulpiride and bromocriptine, but also their interaction effect. If the effects of bromocriptine are mediated by dopaminergic D2 receptors, those should be abolished by co-administration with sulpiride.

Table 1  Factorial pharmacological design. We employed a 2x2 factorial pharmacological design, where pharmacological factors were bromocriptine, a dopamine receptor agonist, and sulpiride, a dopamine receptor antagonist. Each of these factors could be ‘ON’ (drug) or ‘OFF’ (placebo).

<table>
<thead>
<tr>
<th>Session name</th>
<th>Placebo</th>
<th>Bromocriptine</th>
<th>Sulpiride</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1 (Bromocriptine)</td>
<td>OFF</td>
<td>ON</td>
<td>OFF</td>
<td>ON</td>
</tr>
<tr>
<td>Factor 2 (Sulpiride)</td>
<td>OFF</td>
<td>OFF</td>
<td>ON</td>
<td>ON</td>
</tr>
</tbody>
</table>

The dose selection of sulpiride (Dogmatil®, Sanofi-aventis, 400 mg) and bromocriptine (Parlodel®, Novartis, 1.25 mg) was based on previous studies revealing good tolerance (Mehta et al., 2004; Cools et al., 2009; Dodds et al., 2009). Participants received the capsule corresponding to the bromocriptine factor 30 minutes after receiving the one corresponding to the sulpiride factor. The order of drug administration was pseudorandomly assigned and counterbalanced across participants. The resting-state functional resonance imaging (fMRI) started approximately 2 hours after first drug intake and took 7.5 minutes. Participants were instructed to relax and keep their eyes open. The resting-state fMRI data reported here were acquired prior to the acquisition of a task-related fMRI session, during which participants completed a reversal learning task reported in van der Schaaf et al. (van der Schaaf et al., 2014).
The mean time to maximal plasma concentration of sulpiride and bromocriptine is approximately 3 and 2.5 hours, respectively, with a plasma half-life of approximately 12 and 7 hours, respectively (Deleu et al., 2002; Mehta et al., 2004). Accordingly, drug had maximum effects during testing. Subjective mood ratings were measured with the Bond and Lader visual analog scales (Bond and Lader, 1974). Mood measures, blood pressure, and heart rate were taken approximately 30 minutes before, approximately 2 hours after, and approximately 6 hours after first drug intake (reported in van der Schaaf et al. (van der Schaaf et al., 2014)). Participants’ general cognitive performance and mood were not different following different drug sessions indicating that effects were not due to nonspecific drug effects on mood and global cognitive performance.

**Image acquisition and preprocessing**

Structural (T1-weighted MP-RAGE sequence, time echo/time repetition (TE/TR) = 3.03/2300 ms, flip angle = 8°, field of view (FOV) = 256 \times 256 \times 192 mm, voxel size = 1 mm isotropic, GRAPPA acceleration factor 2) and functional images (whole-brain gradient echo planar imaging sequence; TE/TR = 30/1680 ms, flip angle = 70°, FOV = 224 \times 224 \times 137 mm, 39 ascending transverse slices; voxel size = 3.5 \times 3.5 \times 3.0 mm) were collected using a 3-T Siemens MRI scanner with an 8-channel head coil. To reduce signal drop-out and geometric distortions, we used a short TE and reduced echo train length by means of factor 2 accelerated GRAPPA (Griswold et al., 2002).

Images were preprocessed using SPM8 (Wellcome Trust Center for Neuroimaging, London, United Kingdom) and MATLAB. The images were realigned, slice-time corrected and co-registered to the structural image. Participants’ head motion during fMRI acquisition did not differ between experimental sessions, as indexed by the session-specific average head translation and tested with a 2-by-2 full-factorial ANOVA (factors: sulpiride, bromocriptine; all p > 0.05). The images were then smoothed with an isotropic 5 mm full-width half maximum Gaussian kernel. Images were low-pass filtered using a fifth order Butterworth filter to retain frequencies below 0.1 Hz, because the correlations between intrinsic fluctuations are specific to this frequency range (Biswal et al., 1995; Fox and Raichle, 2007). The images were also high-pass filtered (0.008 Hz) to remove low-frequency confounds. To remove non-neuronal fluctuations from the data, we regressed out 27 regressors from each timeseries: 3 regressors describing timeseries of average signal intensity in white-matter, cerebrospinal fluid and in a blank portion of the MR images (out of brain signal) (Helmich et al., 2010); 24 regressors describing timeseries of head motion, namely linear and quadratic effects of the 6 parameters describing the motion of each fMRI image, as well as the first derivative of those effects (to control for spin-history effects) (Lund et al., 2005).
The striatal data were extracted using a striatal mask based on Harvard-Oxford atlas (Flitney et al., 2007). Non-neuronal fluctuations that might be introduced to the data due to individual differences in the striatal size were further controlled using linear regression. Thus, for each subject, we considered the first principal component across voxels within the striatal mask that fell into either white matter or cerebral-spinal fluid (with probability >.99) and regressed out that signal, its square and its cube from all striatal voxels signal.

**Functional parcellation of the striatum**

To ensure a functionally informed parcellation of the striatum, we based our segregation on the functional time-series, using clustering analysis of the correlations among voxel time-series. Furthermore, to ensure that this parcellation scheme was valid at the between-subject level, we performed a stability analysis to identify clusters that were conserved over subjects. We used K-means clustering algorithm to identify different subdivisions of the striatum. In this algorithm, those voxels with higher similarity in correlation pattern of their time-series are more likely to be clustered together. The correlation pattern of each voxel in the striatum was quantified based on its correlation with all other striatal voxels. Therefore, first the correlation matrix for each participant was computed using correlation between each striatal voxel with all other striatal voxels. Next, to compute correlation matrix across group, the individual correlation matrices were Fisher-transformed, averaged and transformed back to correlation space by inverse-Fisher transform. We then used a standard K-means clustering algorithm, using correlation as distance measure, as implemented in MATLAB `kmeans` routine (Mathwork) to parcellate the striatum. Each clustering analysis was replicated 20 times with random initial centroids to avoid local extrema.

K-means clustering operates on a user-defined number of clusters. Since this number is unknown, we performed a stability analysis to identify the most consistent and coherent number of clusters. Subjects were randomly divided into two groups and a series of parcellation into 2 to 8 clusters was carried out separately for each group. The clustering solutions based on data of two groups were then assessed to examine whether they are matched (see Supplementary Appendix for mathematical definition). This procedure was repeated for 100 randomly division of subjects to two groups and used to perform a Monte Carlo randomization test to obtain the largest K resulting in stable clustering solution across group.

This analysis ensures us that this parcellation scheme was valid across the group and results in clusters that could be reliably identified over group. It is important to realize that the goal of functional parcellation and the stability analysis were
not to determine the number of striatal subregions. Rather, the goal of this stability analysis is to find subregions that 1) are consistent at the group level given limitations of fMRI signals; and 2) have distinct pattern of connectivity from the point of view of data (Neubert et al., 2014).

Having established \( K \), we again performed clustering to define the striatal clusters. For every participant and every session, the clustering algorithm defined five clusters according to striatal connectivity matrix of all other remaining subjects in the same session (leave-one-subject-out procedure). The clusters were matched very closely across different sessions and across the 25 cross-validation folds. The leave-one-subject-out procedure ensures that there is no selection bias in definition of regions of interest.

Following classical models of the striatum and to limit model-space for DCM analysis, we considered a 3-nodes architecture for the striatum. Thus, the ventral striatum, dorsal caudate nucleus and dorsal-anterior putamen clusters out of the clustering solution with \( K=5 \) were then chosen as representative of motivational, cognitive and premotor striatum, respectively. For each participant, the spatial intersection of these three clusters across four sessions were generated and used as volume of interests. The first eigenvariate of data in each volume of interest was then extracted for every session.

**Dynamic causal modeling**

We used DCM software implemented in SPM12b (version: 5616). All models were inverted using generalized filtering (Friston et al., 2010; Li et al., 2011a) successfully. The inversion scheme estimated model evidence and fixed-connections and dopaminergic-modulatory parameters (as well as hemodynamic parameters) for each model. The estimated model evidence reflects the plausibility of the model taking into account both goodness of fit and model complexity. We used random-effects Bayesian model comparison to evaluate the plausibility of every model of the model space across the population (Stephan et al., 2009) and report the results in terms of protected exceedance probabilities (Rigoux et al., 2014) throughout the paper.

**Results**

Resting state data were analyzed from twenty five healthy volunteers who participated in a fMRI experiment, in which both resting state as well as task-related data were collected (van der Schaaf et al., 2014). We employed a 2x2 factorial pharmacological design (Table 1). Therefore, each participant was tested
on each cell of this factorial design, receiving 2 different opaque gelatin capsules on each of the four testing sessions, corresponding to the combinations of the two pharmacological factors. Trait impulsivity was indexed with the Barratt Impulsiveness Scale (BIS) (Patton et al., 1995) (Table S1). Previous work with [$^{11}$C] raclopride Positron emission tomography (PET) in healthy volunteers has shown that subjects with high BIS scores exhibit lower D2-receptor availability than do subjects with low BIS scores (Buckholtz et al., 2010). Moreover, BIS scores have been shown to predict the direction of bromocriptine’s effects on striatal BOLD signal (Cools et al., 2007b). The BIS was administered in each session approximately 5.5 hours after first drug intake. The average of the BIS scores across all four sessions was used as an index of trait impulsivity\(^1\), given the high inter-session correlation between those scores (all pair-wise correlations>9), and the absence of significant drug effects on total BIS scores (p>.05 for main effects and interaction, controlled for order effect).

**Defining motivational, cognitive and motor striatal nodes**

There are no reliable in-vivo structural markers of the boundaries between functionally distinct regions of the human striatum (Voorn et al., 2004), namely the motivational, cognitive and motor regions. Here, we overcome this obstacle by using an unsupervised parcellation scheme based on correlation between functional time-series. The parcellation scheme identified 5 clusters reliably at the population-level (p<0.05, Monte-Carlo randomization test, see Methods and Supplementary Appendix for description, Figure S1A). The clustering solution included a ventral striatal region (including nucleus accumbens, ventral caudate nucleus and ventral parts of the putamen), a medial caudate region, a dorsal caudate nucleus region, a dorsal-anterior putamen region and a dorsal posterior putamen region (Figure S1B). The macroanatomical borders of these clusters were consistent with connectivity pattern of striatum measured with various techniques in different species (Haber et al., 2000; Ikemoto, 2007; Draganski et al., 2008b). Although the clustering algorithm was blind to voxel location, there was a very high symmetry between two hemispheres as more than 95 percent of symmetric voxels assigned to the same clusters.

Following classical models of the striatum based on hypothesized functions of striatal regions and its cortical connectivity signature (Alexander et al., 1986; Haber et al., 2000) and to limit model-space for DCM analysis, we considered a 3-nodes architecture for the striatum. Thus, the ventral striatum, dorsal caudate

---

\(^1\) For one subject, BIS scores were obtained only in two out of the four sessions. For this subject, the average across these two sessions was used.
**Figure 1** VOI definition, model definition and model selection. A) Striatal clusters obtained using data-driven parcellation of the human striatum in motivational (ventral striatum, VS, in red), cognitive (dorsal caudate nucleus, DCN, in green) and motor (dorsal-anterior putamen, DAP, in blue) regions. B) Model space of pharmacological input, representing different scenarios for the effects of sulpiride and bromocriptine. Dopaminergic drugs have been included as extrinsic modulatory pharmacological inputs (PIs). These scenarios differ in how they might capture the effects of bromocriptine and sulpiride, namely as only an effect of bromocriptine (PI1) or only an effect of sulpiride (PI2), independent effects of both (PI3), independent effects of both but a potentially nonlinear (and independently estimated) effect of combined administration (PI4), or antagonistic and symmetric effects of both (PI5). Thus, the number of inputs vary across the different sets of PIs. For example, while PI4 contains three inputs (U1, U2 and U3), PI5 contains only one input (U1). In PI5, the only input, U1, is +1 in the bromocriptine session, -1 in the sulpiride session and zero in the placebo and combined session. Therefore, this pharmacological input refers to a situation in which bromocriptine and sulpiride show opponent and symmetric effects, such that co-administration of bromocriptine with sulpiride abolishes the effects evoked when administered alone. C) Models of intra-striatal connectivity that differed in terms of the number and directionality of the connections between the three striatal regions (A-matrix). Colors are associated with different intra-striatal connections. The table highlights which connections are included in each model. For example, A2 includes four connections represented with different colors: magenta (VS→DCN), cyan (DCN→DAP), yellow (DCN→VS) and salmon (DAP→DCN). D) Models of dopaminergic modulatory effects on striatal connections. Colors are associated with different dopaminergic modulatory
nucleus and dorsal-anterior putamen cluster were chosen as representative of motivational, cognitive and premotor striatum, respectively (Figure 1A). We focused on the dorsal caudate nucleus, given its strong associations with cognitive control and associated cortical regions (e.g. the dorsolateral prefrontal cortex (Haber et al., 2000; Draganski et al., 2008b)). We focused on the dorsal-anterior putamen because it is known to be a target of dopaminergic-mediated connectivity from the dorsal caudate nucleus (Haber et al., 2000), while being strongly associated with premotor cortex (e.g. the rostral cingulate motor areas) as well as lateral prefrontal cortex (Calzavara et al., 2007; Draganski et al., 2008b; Helmich et al., 2010). Note that although the rostral cingulate motor area is implicated in premotor functions (e.g. by sending direct projections to the spinal cord (He et al., 1995) and primary motor cortex (Dum and Strick, 2002)), this area is also associated with negative affect, pain and cognitive control (Shackman et al., 2011).

For each participant, the spatial intersection of these three clusters across four sessions were generated and used as volume of interests. The first eigenvariate of data in each volume of interest was then extracted for every session.
**Dopaminergic drug effects on intra-striatal effective connectivity**

We constructed stochastic DCMs to assess dopaminergic drug effects on intra-striatal connectivity using the first eigenvariate of each of the three striatal regions as a summary time-series, after removal of nuisance-related variance.

DCM enjoys a property of Bayesian schemes, namely the ability to dissociate between the goodness of a particular model architecture based on the data, and the consistency (nonzero) of experimental effects on the model parameters across the population. This property is important for the purpose of this study, given that dopaminergic drug effects on the striatum likely vary widely across participants (Cools and D’Esposito, 2011). Accordingly, across the whole group, the dopaminergic drug effects across the group as a whole on intra-striatal connectivity might average around zero. Yet, by using DCM, we could assess the degree to which the drug alters intra-striatal connectivity even when the sign of this effect differs across participants. Specifically, we included dopaminergic drugs as modulatory pharmacological inputs in DCM, and then used Bayesian model comparison to assess effects of sulpiride and bromocriptine across the group, independent of the sign of their effects between individuals.

**Model space**

Our model space had three factors. These included the nature of the drug effects (i.e. the pharmacological inputs, see Fig 1B), the underlying intra-striatal connectivity architecture (Fig 1C), and the modulation of these intra-striatal connections by dopamine (Fig 1D). The first factor concerned the effects of the pharmacological manipulation, modeled as extrinsic modulatory pharmacological input, enabling us to assess both the main and interaction effects of our two pharmacological factors sulpiride and bromocriptine. This resulted in five sets of pharmacological input (PI1 to PI5 in Fig 1B). The second factor concerned the presence of directed connections among the three striatal nodes, yielding in total five types of hypothetical striatal architecture (A1 to A5 in Fig 1C). Finally, the effects of dopamine on these five types of intra-striatal architecture were investigated by allowing pharmacological (dopaminergic) inputs to modulate the intra-striatal architectures in five different ways (B1 to B5 in Fig 1D), in addition to a null model with no modulation (B0 in Fig 1D). It should be noted that our model space did not include models in which dopamine was allowed to modulate non-existent connections, as this was considered biologically implausible (e.g. A1B2). In total, we considered 15 different mesostriatal architectures (see the list in Fig 1F). Here, 5 models were null models with no modulation but with different underlying architectures (i.e. the B0 models). The combination of the other 10 mesostriatal architectures (B1 to B5) with the 5 sets of pharmacological inputs...
(PI1 to PI5) resulted in a total of 50 models. Together with the 5 null models, this resulted in a final model space of 55 models.

The first factor concerned different scenarios of the effects of pharmacological drugs: (i) A main effect of bromocriptine, but no effect of sulpiride. In this model, bromocriptine, but not sulpiride, was allowed to change striatal connectivity (Figure 1B, PI1); (ii) A main effect of sulpiride, but no effect of bromocriptine. In this model, sulpiride, but not bromocriptine, was allowed to change striatal connectivity (Figure 1B, PI2); (iii) A main effect of sulpiride, a main effect of bromocriptine and no interaction between sulpiride and bromocriptine. In this model, both sulpiride and bromocriptine were allowed to change the striatal connectivity independently, and the effect of the combined session corresponded to the sum of their effect when administered alone (Figure 1B, PI3); (iv) A main effect of sulpiride, a main effect of bromocriptine and an interaction effect of sulpiride and bromocriptine. In this model, both sulpiride and bromocriptine were allowed to change striatal connectivity independently as well as their interaction. Namely, in the combined session, the effects of sulpiride and bromocriptine could vary independently of their effects when administered alone (Figure 1B, PI4). (v) Symmetric effects of sulpiride and bromocriptine on striatal connectivity (Figure 1B, PI5). In this model, the magnitude of the effect of bromocriptine was equal to that of sulpiride, but in the opposite direction.

The second factor concerned intra-striatal connections, independently from the dopaminergic modulations (Figure 1C). We created 4 models containing forward connections from the ventral striatum to the dorsal caudate nucleus and from the dorsal caudate nucleus to the dorsal-anterior putamen. This feature of the models is grounded in neuroanatomical evidence from non-human primates that demonstrate the presence of forward information flow along the mediolateral gradient across the striatum (Haber et al., 2000). This property, as well as contribution of backwards connections, was assessed by constructing four intra-striatal architectures: i) a forward connection from the ventral striatum to the dorsal caudate nucleus and from the dorsal caudate nucleus to the dorsal-anterior putamen (Figure 1C, A1); ii) both forward and backward connections between the ventral striatum and the dorsal caudate nucleus and between the dorsal caudate nucleus and the dorsal-anterior putamen (Figure 1C, A2); iii) forward connections from the ventral striatum to the dorsal caudate nucleus, from the dorsal caudate nucleus to the dorsal-anterior putamen and from the ventral striatum to the dorsal-anterior putamen (Figure 1C, A3); iv) forward and backward connections between all three subregions (Figure 1C, A4). Finally, we included a model with v) two connections, one from the ventral striatum to the dorsal caudate nucleus and the other one from the ventral striatum...
to the dorsal-anterior putamen. This model was created based on data showing that the ventral striatum sends tri-synaptic projections to the primary motor cortex and to prefrontal cortex (Kelly and Strick, 2004), which could result in modulation of the dorsal caudate nucleus and dorsal-anterior putamen by modulating their associated cortical areas in the cognitive and motor loops of frontostriatal circuitry.

The third factor concerned the effects of dopamine on intra-striatal connectivity. We constructed models that allowed modulatory effects of dopamine on all (Figure 1D, B4), some (Figure 1D, B1, B2, B3 and B5) or none (Figure 1D, B0) of the striatal architectures described above. The effects of sulpiride and bromocriptine were assumed to be homogeneous with respect to input type, across the different striatal connections.

Hypotheses
Based on neuroanatomical and neurochemical evidence (Haber et al., 2000; Ikeda et al., 2013), we hypothesized that dopaminergic drugs would modulate the flow of information in a directional forward fashion along the mediolateral gradient in the striatum. Furthermore, previous PET work has shown that impulsivity-dependent dopaminergic effects are mediated by the D2 receptor (Dalley et al., 2007; Buckholtz et al., 2010). Accordingly, we anticipated that our data would be best fit by model A1B1 and that individual differences in dopaminergic drug effects, as indexed by the modulatory B parameters, would depend on trait impulsivity. To test this hypothesis we assessed not only the effects of the D2-receptor agonist bromocriptine, which also has affinity for the D1-receptor (while also altering noradrenalin transmission), but also the effects of sulpiride, a highly selective antagonist for the D2-receptor. In addition, we assessed in a combined session whether the effects of bromocriptine would be blocked by pretreatment of sulpiride. We predicted that the effect of bromocriptine would be opposite to that of sulpiride, and that these would not interact. If the effect of bromocriptine would be equal in size to that of sulpiride, then the combined administration would be indistinguishable from that of placebo. In this case, the data would be best fit by input set PI5. However, if the effect of bromocriptine and sulpiride are independent but of unequal size, then the data would be best fit by input set PI3.

Model selection
We employed a two-step model selection approach. First, we performed a family-wise random-effect Bayesian model comparison to test the pharmacological drug effects (Figure 1E) (Penny et al., 2010). Second, we performed random-effect Bayesian model comparison to compare different mesostriatal architectures given the winner input in the previous step (Figure 1F) (Rigoux et al., 2014).
First, Bayesian model comparison over the model space of pharmacological input revealed very strong evidence in favor of the pharmacological input family with opponent, symmetric effects of sulpiride and bromocriptine (P15, protected exceedance probability of 1.00, expected posterior model probability of 0.50; Figure 1E). Thus, sulpiride and bromocriptine altered intra-striatal connectivity to the same degree, but in opposite directions, consistent with the hypothesis that dopaminergic drug effects on mesostriatal connectivity are mediated by the D2-receptor. In this winning model, the effects of sulpiride and bromocriptine on intra-striatal coupling cancelled each other out, leaving a zero net effect.

Second, Bayesian model comparison over the model space of mesostriatal architectures revealed evidence in favor of the architecture with forward and backward projections between the ventral striatum and the dorsal caudate nucleus as well as between the dorsal caudate nucleus and the dorsal-anterior putamen (A2B2, protected exceedance probability of 0.92, expected posterior model probability of 0.40; Figure 1F). In summary, the winning model contains bidirectional connections between striatal regions, in a hierarchical fashion, and a modulatory input on each of these connections to model the opponent dopaminergic drug effects.

These results were robust to model selection procedures: A one-step random-effect Bayesian model comparison among all 55 models revealed that the same model, A2B2, best explained our data across the whole model space (protected exceedance probability of 0.95, expected posterior model probability of 0.19).

**Trait impulsivity and dopaminergic drug effects on intra-striatal effective connectivity**

Further analysis of the characteristics of the winning model led to an important additional insight on how dopamine modulates intra-striatal connectivity. First, using one-sample t-tests, we confirmed that each of the 4 parameters quantifying intra-striatal connectivity in the winning model were significantly above zero (all p <.006, Bonferroni-corrected, Table S2). This observation is consistent with the expectation that the three striatal subregions are strongly connected. Second, the same statistical procedure revealed that none of the 4 parameters quantifying dopaminergic modulation of intra-striatal connectivity was significantly different from zero (all p>0.006, Bonferroni-corrected, Table S2). This null-effect persisted even when the statistical threshold was relaxed to 0.05 uncorrected for multiple comparisons (Table S2). This null-effect might seem to contradict the model-selection results. In fact, these observations indicate that dopaminergic modulatory inputs explain significant variance in the fMRI timeseries, despite the fact that the sign of the modulatory effect is inconsistent across subjects. The latter finding fits
with the known inter-individual variability in the direction and extent of dopaminergic drug effects (Cools and D’Esposito, 2011). Strong dopaminergic drug effects in individual participants often add to zero when averaged across a group (Cools and D’Esposito, 2011). Given that individual differences in dopaminergic drug effects on striatal activity have been shown to depend on trait impulsivity according to D2-receptor density (Cools et al., 2007b; Dalley et al., 2007; Buckholtz et al., 2010), we hypothesized that effects of D2-receptor agents on intra-striatal connectivity is associated with trait impulsivity. Therefore, we tested whether modulatory input parameters of the winning model are associated with trait impulsivity. This analysis was implemented through a repeated measures ANOVA, assessing individual modulatory parameters as a function of connection direction (forward versus backward), striatal pair (ventral striatum-dorsal caudate nucleus or dorsal caudate nucleus- dorsal-anterior putamen) and trait impulsivity (BIS scores). This analysis revealed a significant positive association between impulsivity and connection strength ($F(1,23)=4.52, p=0.044$). Crucially, there was a significant three-way interaction between trait impulsivity, striatal pair, and connection direction ($F(1,23)=4.71, p = .041$). Post-hoc correlation analyses revealed that the three-way interaction was due to a highly significant positive correlation between trait impulsivity and the drug effects on the forward connection from the ventral striatum to the dorsal caudate nucleus ($r = 0.53, p= 0.007$, Figure 2B). Trait impulsivity did not correlate with drug effects on the other connections ($p > 0.05$, Table S3). These results indicate that trait impulsivity is associated with the increasing and decreasing effects of respectively bromocriptine and sulpiride on the forward connection from the ventral striatum to the dorsal caudate nucleus, but not on the other connections. Thus, trait impulsivity is associated with increased dorsal caudate nucleus drive from ventral striatum by stimulation of D2-receptors, and a decreased dorsal caudate nucleus drive from ventral striatum by blockade of D2 receptors.

We also performed two control analysis regarding association of impulsivity with connectivity between ventral striatum and dorsal caudate nucleus. First, we conducted a control analysis using BIS scores in the placebo session, instead of the original analysis with mean across all four sessions, as the index of trait impulsivity. The results of this control analysis were consistent with those of main analysis (SI text). Second, we conducted a relatively model-free analysis (linear regression) to confirm our findings regarding association of impulsivity with the connectivity between ventral striatum and dorsal caudate nucleus, independent of the estimated effective connectivity strengths in the winning mesostriatal architecture from DCM (SI text). The results of this analysis were consistent with those found based on DCM (Figure S2, Table S4).
Dopamine Modulates the Human Striatum Functional Architecture

It is known that selection of the best model among large number of competing hypotheses could be fragile, especially if the data of different subjects could be fitted by different models (as could happen in a random effect model space) (Penny et al., 2010). Therefore, we have tested our a priori hypotheses on striatal connectivity on models comprising 3 nodes. This approach generates robust inferences, but it also limits the inference of the study to ventral and dorsal-anterior portions of the striatum. Here, following a reviewer’s comment, we build on those findings and extend the analysis to a more posterior part of the striatum. Connectivity-based parcellation of the striatum identified a dorsal posterior putamen cluster (Figure 3A), known to be connected to motor cortex and strongly implicated in motor control (Draganski et al., 2008b; Helmich et al., 2010) and habitual action selection (Wunderlich et al., 2012a). Neuroanatomical evidence in

**Figure 2** The relationship between dopaminergic modulation of striatal ventrodorsal connectivity and trait impulsivity. Scatter-plot shows the relationship between trait impulsivity and the modulatory parameter encoding dopaminergic modulation of the connection from ventral striatum to dorsal caudate nucleus in the winning model. Trait impulsivity is associated with drug-induced increases (decreases) by D2 dopamine agonist (D2 dopamine antagonist) of dorsal caudate nucleus input from VS. Abbreviations: BIS, Barratt impulsiveness score; VS, ventral striatum; DCN, dorsal caudate nucleus; Hz, Hertz.
non-human primates (Haber et al., 2000; Ikeda et al., 2013) suggests that dopamine modulates forward connections along the mediolateral gradient across the striatum. Therefore, we extended the intra-striatal architecture by adding forward connection from dorsal-anterior putamen to dorsal-posterior putamen (Figure 3B, a1). Two models of dopaminergic modulation associated with this architecture were tested, where dopamine either modulated or did not modulate this connection (Figure 3B, b1 and b0, respectively). A second intra-striatal architecture was created by including bidirectional connections between dorsal-anterior putamen and dorsal-posterior putamen. Three models of dopaminergic modulation associated with this architecture were tested, where dopamine modulated none (Figure 3B, b0), the forward connection (Figure 3B, b1) or both connections (Figure 3B, b2).

**Figure 3** Post-hoc analysis of dopamine-mediated connectivity of dorsal-posterior putamen (DPP). A) DPP cluster (in brown) obtained using data-driven parcellation of the human striatum. B) Three models of intra-striatal connectivity of DPP and four models of its dopaminergic modulation were created. The table highlights which connections are included in each model. In total, 7 models were tested. Note that the pharmacological input as well as mesostriatal connections among other striatal regions are fixed according to the optimal model presented in Figure 2. C) Random-effect Bayesian model comparison for 7 mesostriatal architectures with DPP. Bayesian model comparison strongly favored a2b2 among all 7 models. D) The winning mesostriatal architecture. Striatal areas are bidirectionally connected along a mediolateral gradient and dopamine modulates adjacent areas along this gradient. Abreviations: VS, ventral striatum; DCN, dorsal caudate nucleus; DAP, dorsal anterior putamen; DPP, dorsal posterior putamen; DA, dopamine; BMC, Bayesian model comparison.
from dorsal-anterior putamen to dorsal-posterior putamen. Finally, we included a third intra-striatal architecture where ventral striatum directly modulated dorsal-posterior putamen (Figure 3B, a3). This model was created based on data showing that the ventral striatum sends tri-synaptic projections to the primary motor cortex and to prefrontal cortex (Kelly and Strick, 2004), which could result in a ventral-striatal modulation of dorsal-posterior putamen through the motor loop of the frontostriatal circuitry. Two models of dopaminergic modulation associated with this architecture were tested, where dopamine either modulated (Figure 3B, b3) or did not modulate this connection (Figure 3B, b0).

These models were fitted and compared using random-effects Bayesian model comparison. This analysis revealed strong evidence in favor of the mesostriatal architecture with forward and backward baseline connections between the dorsal-anterior putamen and the dorsal-posterior putamen, where dopamine modulated both connections (Figure 3C, protected exceedance probability of 1.00, expected posterior model probability of 0.81). These findings extend our prior findings to more posterior parts of the striatum. The findings suggest that dopamine modulates both forward and backward intra-striatal connections along the mediolateral axis of the striatum (Figure 3D).

**Discussion**

This pharmacological-fMRI study addresses the functional architecture of the human striatum and dopaminergic influences on striatal information processing (Cools et al., 2007b; Dalley et al., 2007; Belin et al., 2008). We manipulated the connectivity between motivational, cognitive, and motor portions of the striatum with dopaminergic drugs, and we exploited inter-individual differences in mesostriatal dopamine systems to explain trait-dependent effects of the dopaminergic manipulations. Striatal connectivity patterns were quantified with stochastic dynamic causal modeling of intrinsic BOLD activity measured in a within-subject, double-dummy, placebo-controlled cross-over design. There are two main findings. First, Bayesian model comparison indicates that human striatal architecture is sparse and largely consistent with neuroanatomical data from non-human primates and rodents (Haber et al., 2000). Namely, functional interactions between the ventral striatum and the dorsal-anterior putamen are mediated by the dorsal caudate nucleus, and the efficacy of those interactions is modulated by dopaminergic tone. Second, the magnitude of the dopaminergic modulation of a portion of those interactions depends on trait-impulsivity. Namely, highly impulsive individuals have increased sensitivity to dopamine-induced
changes in information flow from the ventral to the dorsomedial striatum. This result might explain how, in highly impulsive individuals, cognitive processes supported by the dorsomedial striatum can become particularly vulnerable to the motivational drive from the ventral striatum (Lawrence and Brooks, 2014).

**Intrinsic striatal architecture**
This study shows that a model of striatal connectivity without a direct connection between the ventral striatum and the dorsal-anterior putamen fitted the data significantly better than models with such a connection. This finding suggests that communication between those two striatal regions is mediated by the dorsal caudate nucleus. In macaques, dopamine mediates information flow along the mediolateral pathway through serial reciprocal connections between the striatum and the midbrain (Haber et al., 2000). Accordingly, we interpret the effects of stimulation and blockade of D2-receptors at the level of the striatum in terms of altered midbrain-mediated feedforward information flow from the ventral striatum to the dorsal caudate nucleus. However, we cannot exclude concurrent actions via modulation of topographically specific, feedforward circuits connecting the prefrontal cortex with the striatum (McFarland and Haber, 2002; Honey et al., 2003; Haber and Knutson, 2010; Cole et al., 2013a, 2013b).

**Impulsivity amplifies dopaminergic modulations of ventrodorsal striatal connectivity**
Work with behaving rodents indicates that the transition from impulsive to compulsive drug use, and the corresponding transition of behavioural control from the ventral to the dorsolateral striatum, can be promoted by dopamine (Dalley et al., 2007; Belin and Everitt, 2008; Belin et al., 2008). Dalley et al. (Dalley et al., 2007) have shown that ventral striatal D2 receptors density predicts individual difference in trait impulsivity, which itself predict propensity to cocaine seeking (Dalley et al., 2007) and addiction-like behaviour (Belin et al., 2008). Our findings add to this body of knowledge by showing that trait impulsivity is associated with the influences of striatal D2 receptors on how effectively the ventral striatum modulates activity in the dorsal caudate nucleus. This effect was specific: The dopaminergic drugs were found to modulate connectivity along both the ventral striatum to dorsal caudate nucleus and the dorsal caudate nucleus to dorsal-anterior putamen pathway, whereas impulsivity was associated with only the first part of that pathway. Work with experimental animals (Dalley et al., 2011) raises the intriguing possibility that compulsivity might convey vulnerability of the connection between dorsomedial and dorsolateral striatum to dopaminergic drugs.
Limitations
It can be argued that the current findings are statistical constructs of an oversimplified model of intrinsic striatal connectivity. In fact, stochastic DCM provides an objective and quantitative procedure for distinguishing between explicit models of functional anatomy (Friston et al., 2011; Li et al., 2011a; Daunizeau et al., 2012; Kahan et al., 2014). The model space was simplified to the core elements relevant to understand how dopamine modulates intrinsic striatal connectivity. DCM relies on prior assumptions about the distribution of a number of model parameters, but the findings are robust, having being confirmed with an independent model-free analysis.

In this study, we modeled dopaminergic drugs as extrinsic inputs in DCM and demonstrated that these inputs modulate striatal connectivity in a specific topographic fashion. This modulation could be mediated by different elements of mesostriatal circuitry. One possibility is that sulpiride and bromocriptine modulate activity of dopamine cells in the ventral tegmental area (VTA) and substantia nigra pars compacta (SNC). Another possibility is that these drugs modulate D2 receptors in the striatum directly. In theory, it is possible to dissociate between these two possibilities by including VTA and SNC as additional nodes in DCM. In practice, it is not possible to get reliable BOLD signals from these regions with the standard whole-brain fMRI settings used in this study. Namely, the anatomical location of these regions makes their BOLD signals exquisitely sensitive to both physiological artifacts and subject motion. The consequences of those artifacts are particularly deleterious during resting-state fMRI. Therefore, we have not included data from these two regions in this study and focused on their downstream effects on the striatum. Note that even if dopaminergic drug effects are mediated by modulation of dopamine cells in the midbrain, this does not invalidate the current approach. In this case, we can assume that we have modeled midbrain as a hidden node in DCM and instead of fitting models to data from this region, focused on its downstream effects. This approach has been validated previously in the context of DCM for fMRI (David et al., 2011; Moran et al., 2011; Marreiros et al., 2013; Kahan et al., 2014).

To define regions of interest for DCM, the striatum was parcellated into functionally homogeneous regions using a data-driven method. This connectivity-based parcellation approach overcomes the known difficulty of defining boundaries between functionally distinct regions of the human striatum. For instance, the ventral striatal cluster extended beyond the nucleus accumbens, including the ventromedial caudate nucleus and the rostroventral putamen, in line with neurophysiological studies (Voorn et al., 2004). More generally, our data-driven
mediolateral parcellation of the striatum corresponds closely with regions identified on the basis of neurophysiological data, and follows the pattern of excitatory cortical, thalamic and amygdaloid inputs to the striatum (Alexander et al., 1986; Voorn et al., 2004; Draganski et al., 2008b; Haber and Knutson, 2010). We limited the main analyses to three regions that could be linked to the known mediolateral organization of the striatum and that have been previously shown to be implicated in motivational, cognitive and premotor circuits (Alexander et al., 1986; Haber et al., 2000; Draganski et al., 2008b). A restricted model space generates statistically robust inferences (Penny et al., 2010), but it also limits the inference of this study to ventral and dorsal-anterior portions of the striatum. A post-hoc extension of the analyses to motor striatum, the dorsal-posterior putamen, connections provides strong evidence in favor of a mesostriatal architecture with forward and backward baseline connections between the dorsal-anterior putamen and the dorsal-posterior putamen, with dopamine modulating both connections. These data suggest that dopamine modulate both forward and backward intra-striatal connections along the mediolateral axis of the striatum.

We found evidence that dopamine modulates striatal connectivity not only along the ventral to dorsal pathway, but also along the dorsal to ventral pathway. At first glance, this finding cannot be reconciled with the model of nigro-striato-nigral connectivity observed in macaques (Haber et al., 2000), which is known to be unidirectional. However, model connections between two nodes are not limited to anatomical monosynaptic connections. Accordingly, it is possible that the dorsal striatum might affect the ventral striatum through its connection with the prefrontal cortex. There is neuroanatomical evidence in nonhuman primates that motivational and cognitive areas of the striatum show converging cortical inputs (Haber et al., 2006). Another possibility is that dorsal striatum affects ventral and medial striatum indirectly through its connection via the mediodorsal thalamus, which itself projects to the caudate nucleus (McFarland and Haber, 2001, 2002). Indeed, dopaminergic stimulation and blockade of D2 receptors could modulate the thalamus via inhibitory projections from the dorsal striatum to the thalamus, thereby affecting thalamic input of the caudate nucleus.

Conclusion
Building on recent anatomical work (Haber et al., 2000), this study provides empirical evidence for a hierarchical architecture in the flow of information within the human striatum. Communication between the ventral and the dorsal putamen is mediated by the dorsal caudate nucleus. This architecture points to structured interactions between frontostriatal loops that have long been considered to have limited anatomical convergence (Selemon and Goldman-Rakic,
1985; Alexander et al., 1986). Furthermore, this study shows how those interactions are modulated by dopaminergic tone. State-related effects, induced by pharmacological interventions, influenced the striatal circuitry along the mediolateral pathway. These effects are consistent with a midbrain-mediated dopaminergic influence on striatal connectivity (Haber et al., 2000). Trait-related effects, indexed by the interaction between impulsivity and pharmacological interventions, influenced connectivity between the ventral and the dorsal caudate nucleus. This effect is consistent with the notion that impulsivity marks a stronger dopamine-dependent influence of the ventral onto the dorsomedial striatum. One implication of this finding is that, in highly impulsive individuals, early drug-intake episodes could quickly lead to goal-directed drug-intake (Corbit et al., 2012).
Supplementary data

Supplementary Appendix
To ensure that the parcellation scheme is valid at the between-subject level, we performed a stability analysis to identify the largest number of clusters resulting in a clustering solution conserved over subjects. To achieve this, we assessed whether two clustering solutions calculated based on two independent datasets (e.g. by diving subjects randomly to two groups) were matched. Here, we provide a mathematical explanation of our approach.

Two sets of clusters (A and B, each with K clusters) were defined as matched based on the following criteria: First, for every cluster in A and every cluster in B, an overlap index was defined, which corresponds to the number of voxels that overlap between the two clusters. Specifically, for every cluster $a_i$ in A and every cluster $b_j$ in B, the overlap index was defined as $N_{i,j}/\min(N_i,N_j)$, where $N_i$, $N_j$ and $N_{i,j}$ are the number of voxels in $a_i$, $b_j$ and their intersection, respectively. Next, for every cluster $a_i$ in A, $b_j$ in B was defined as matched if it had the largest overlap index with $a_i$. Finally, A and B were considered as matched if each cluster in A was matched with one and only one cluster in B; and vice versa if each cluster in B was matched with one and only one cluster in A. This procedure also gives a one-to-one mapping between “labels” of clusters in A and B, regardless of anatomical location of voxels.

Control analysis with Barratt impulsiveness scale (BIS) in the placebo session as the index of trait impulsivity
We conducted a control analysis using BIS scores in the placebo session (instead of the mean across all four sessions presented in the main text) as the index of trait impulsivity. The results of this control analysis were consistent with those of main analysis.

Similar to the main analysis, a repeated measures ANOVA was employed to assess individual modulatory parameters as a function of connection direction (forward versus backward), striatal pair (ventral striatum and dorsal caudate nucleus or dorsal caudate nucleus and dorsal-anterior putamen) and trait impulsivity (BIS score obtained in the placebo session). Please note that for one of the subjects, BIS was not administered in the placebo session; so data from other subjects were analyzed here. This analysis revealed a significant interaction between trait impulsivity, striatal pair and connection direction ($F(1,22)=4.43$, $p = .047$). Post-hoc correlation analyses revealed that the three-way interaction was due to a highly significant positive correlation between trait impulsivity and the drug effects on the forward connection from the ventral striatum to the dorsal caudate nucleus ($r$
Dopamine Modulates the Human Striatum Functional Architecture

\[ r = 0.51, p = 0.01 \] Trait impulsivity did not correlate with drug effects on the other connections \((p > 0.05)\).

**Model-free analysis of trait impulsivity and dopaminergic drug effects on intra-striatal functional connectivity**

We conducted a control analysis regarding association of impulsivity with functional connectivity between ventral striatum and dorsal caudate nucleus, which did not depend on the mesostriatal architecture selected from DCM analysis.

Similar to the dynamic causal modeling analysis, the first eigenvariates extracted from the ventral striatum, dorsal caudate nucleus and dorsal-anterior putamen clusters (Figure 1A) were used for analysis. These time-series were employed to compute functional connectivity (correlation) between the striatal regions for each session. First, to quantify the strength of the intra-striatal connections, we computed the average connectivity across the four sessions for each subject and each pair of striatal regions. Next, to quantify the dopaminergic drug effect, we computed the difference in functional connectivity (for each pair of striatal regions) between the bromocriptine and sulpiride sessions (bromocriptine minus sulpiride). A regression analysis was conducted with the three connections and the three dopaminergic drug effects (on those connections) as predictors (as well as an intercept) and with trait impulsivity as dependent variable (Figure S2A). This analysis investigates the (partial) correlation between each regressor and trait impulsivity while controlling for the variance explained by the other regressors. It revealed that impulsivity was selectively associated with dopaminergic drug effects on the connectivity between ventral striatum and dorsal caudate nucleus \((p = 0.027, \text{Figure S2B, Table S4})\). The average strength of connections (across drug sessions) did not vary with trait impulsivity, and there was no significant association between impulsivity and drug effects on the two other connections (all \(p > .05, \text{Table S4}\)).
Figure S1 Parcellation of human striatum based on functional connectivity. A) Stability index as a function of number of clusters, K. The 5-clusters solution is the largest K resulting in stable solution across participants. B) The 5-clusters solution shown in several coronal and axial slices. The clustering solution included a ventral striatal region (including nucleus accumbens, ventral caudate nucleus and ventral parts of the putamen; in red), a medial caudate region (in yellow), a dorsal caudate nucleus region (in green), a dorsal-anterior putamen region (in blue) and a dorsal posterior putamen region (in magenta).
Figure S2 The association between trait impulsivity and dopamine-mediated changes in striatal connectivity obtained from the model-free analysis. A) To quantify the strength of the intra-striatal connections, the functional connectivity between the three striatal regions quantified as the mean correlation across all sessions. The difference in functional connectivity (correlation) between bromocriptine and sulpiride session was used as the dopaminergic drug effects on intra-striatal connectivity. A regression analysis with these six regressors (as well as an intercept) conducted to examine their relationship with trait impulsivity. B) Scatter-plot of the relationship between trait impulsivity (BIS scores) and the dopamine-mediated changes on the coupling between ventral striatum and dorsal caudate nucleus. The values in the x-axis are the differences in the connectivity between ventral striatum and dorsal caudate nucleus (bromocriptine minus sulpiride). The values in the x-axis are adjusted for other regressors. Abbreviations: VS, ventral striatum; DCN, dorsal caudate nucleus; DAP, dorsal-anterior putamen; DA, dopamine; BIS, Barratt impulsiveness scale.
### Table S1  Barratt impulsiveness scale (BIS) in each session. Means are shown with standard errors in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>Sulpiride</th>
<th>Bromocriptine</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIS</td>
<td>63.7 (1.9)</td>
<td>62.6 (1.8)</td>
<td>63.8 (1.9)</td>
<td>64.2 (1.9)</td>
</tr>
</tbody>
</table>

### Table S2  Fixed-connection parameters values (A) representing intra-striatal connectivity and dopaminergic-modulatory parameters values (B) of those connections in the winning model, A2B2. Means are shown with standard errors in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>VS-&gt;DCN</th>
<th>DCN-&gt;DAP</th>
<th>DCN-&gt;VS</th>
<th>DAP-&gt;DCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (Hz)</td>
<td>0.1269 (0.018)</td>
<td>0.0820 (0.008)</td>
<td>0.0621 (0.062)</td>
<td>0.1377 (0.138)</td>
</tr>
<tr>
<td>B (Hz)</td>
<td>0.0086 (0.0066)</td>
<td>0.0036 (0.0059)</td>
<td>0.0068 (0.0048)</td>
<td>0.0057 (0.0081)</td>
</tr>
</tbody>
</table>

Abbreviations: VS, ventral striatum; DCN, dorsal caudate nucleus; DAP, dorsal-anterior putamen; DA, dopamine; BIS, Barratt impulsiveness scale.

### Table S3  Correlation of dopaminergic-modulatory parameters values in the winning model with Barratt impulsiveness scale (BIS). Effect size is the r-value.

<table>
<thead>
<tr>
<th></th>
<th>VS-&gt;DCN</th>
<th>DCN-&gt;DAP</th>
<th>DCN-&gt;VS</th>
<th>DAP-&gt;DCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect-size</td>
<td>0.53</td>
<td>0.09</td>
<td>0.37</td>
<td>0.31</td>
</tr>
<tr>
<td>p-value</td>
<td>0.007</td>
<td>0.670</td>
<td>0.067</td>
<td>0.138</td>
</tr>
</tbody>
</table>

### Table S4  Model-free analysis of relationship between trait impulsivity and mesostriatal connectivity. Regression coefficients are shown with standard errors in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>VS-DCN</th>
<th>DCN-DAP</th>
<th>DCN-DAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dopamine-mediated changes in connectivity</td>
<td>21.80 (9.08)*</td>
<td>-14.91 (8.09)</td>
<td>6.86 (13.88)</td>
</tr>
<tr>
<td>Baseline intra-striatal connectivity</td>
<td>0.06 (18.09)</td>
<td>37.98 (21.55)</td>
<td>-14.61 (15.67)</td>
</tr>
</tbody>
</table>
Impulse control disorders in Parkinson’s disease are associated with dysfunction in stimulus valuation but not action valuation.
Abstract

A substantial subset of Parkinson’s disease (PD) patients suffers from impulse control disorders (ICDs), which are side effects of dopaminergic medication. Dopamine plays a key role in reinforcement learning processes. One class of reinforcement learning models, known as the actor-critic model, suggests that two components are involved in these reinforcement learning processes: a critic, which estimates values of stimuli and calculates prediction errors, and an actor, which estimates values of potential actions. To understand the information processing mechanism underlying impulsive behavior, we investigated stimulus and action value learning from reward and punishment in four groups of participants: on-medication PD patients with ICD, on-medication PD patients without ICD, off-medication PD patients without ICD, and healthy controls. Analysis of responses suggested that participants used an actor-critic learning strategy and computed prediction errors based on stimulus values rather than action values. Quantitative model fits also revealed that an actor-critic model of the basal ganglia with different learning rates for positive and negative prediction errors best matched the choice data. Moreover, whereas ICDs were associated with model parameters related to stimulus valuation (critic), PD was associated with parameters related to action valuation (actor). Specifically, PD patients with ICD exhibited lower learning from negative prediction errors in the critic, resulting in an underestimation of adverse consequences associated with stimuli. These findings offer a specific neurocomputational account of the nature of compulsive behaviors induced by dopaminergic drugs.
Introduction

Dopaminergic medications, especially D2 agonist drugs, trigger impulse control disorders (ICDs) such as hypersexuality, binge eating, and pathological gambling in a subset of Parkinson’s disease (PD) patients (Voon et al., 2007). Although PD is primarily associated with dopamine depletion in the substantia nigra and dorsal striatum (Kish et al., 1988), the underlying neural substrates of ICD in PD are mostly the ventral regions of the striatum and their dopaminergic innervations from the ventral tegmental area (Dagher and Robbins, 2009; Voon et al., 2010). Therefore, dopamine neurons projecting to the ventral striatum are relatively intact in PD patients (Kish et al., 1988). Furthermore, it has been suggested that the restoration of dopamine transmission in the dorsal striatum may lead to excessive dopamine receptor stimulation in the ventral striatum (Swainson et al., 2000; Cools et al., 2001), thus inducing ICD in some patients (Cools et al., 2003; Dagher and Robbins, 2009).

Overwhelming evidence has shown that dopamine neurons encode prediction error (PE) signaling, which guides stimulus and action value learning in reinforcement learning (RL) models (Schultz et al., 1997; Bayer and Glimcher, 2005; Pessiglione et al., 2006). It has also been shown that a popular RL model, known as Q-learning (QL), is useful for understanding the mechanistic differences in learning between on- and off-medication PD patients (Frank et al., 2007a; Rutledge et al., 2009). Although it has been hypothesized that the functional dissociation of striatal subregions is critical to understanding the underlying mechanism of compulsive behaviors in both the general population (Everitt and Robbins, 2005; Belin et al., 2013) and PD patients (Cools et al., 2007a; Dagher and Robbins, 2009), previous RL models of PD have not addressed the different roles of the ventral and dorsal striatum in the development of ICD in PD. A well known RL framework that models the different roles of the dorsal (motor) and ventral (limbic) striatum is the actor-critic (AC) framework (Barto, 1995; Dayan and Balleine, 2002). This framework has two modules, known as the critic and the actor, where the former is responsible for PE computations and stimulus value learning and the latter is responsible for action valuation and selection. Empirical studies suggest that the ventral and the dorsal striatum play different roles in decision making, with the former corresponding to the critic and the latter corresponding to the actor (Cardinal et al., 2002; Packard and Knowlton, 2002; O’Doherty et al., 2004). Based on these neuroanatomical data and a prior AC model of addiction (Piray et al., 2010), we here hypothesize that, whereas PD is associated with the actor (i.e., action valuation and selection), ICDs in PD are associated with the critic (i.e., stimulus valuation and PE computations). Therefore, we provide a novel modeling
approach that combines the concept of separate roles for positive and negative PEs in learning (Frank et al., 2007a) with the AC framework to test this hypothesis.

Methods

Participants
This study was part of a larger project conducted at Ain Shams University Hospital, Cairo, Egypt. Participants were asked whether they were willing to participate in the short or long version of the project. In the short version, participants completed 80 trials of a probabilistic learning task compared to 160 trials for the long version. 95 participants were recruited, 79 of which participated in the long version of the project. For this report, we only included those subjects that participated in the long version of the task (with 160 trials). We did not include the data of subjects who participated in the short version of the task in order to have the same number of data points across all subjects for estimating the parameters of computational models. This is because, in principle, (within subject) variance of parameters estimated based on 80 trials is larger than those estimated based on 160 trials and this could inflate statistical comparisons between groups.

Data from 3 participants was discarded from the analysis because these participants had failed to respond in at least 20% of trials. Thus, 4 groups were included in the analyses: 1) PD patients without ICD tested off medication (PD-OFF, n=25, 6 females); 2) PD patients without ICD tested on medication (PD-ON, n=15, 3 females); 3) PD patients with ICD tested on medication (PD-ON-ICD, n=16, 2 females) and 4) healthy controls (n = 20, 7 females). The healthy control participants did not have any history of neurological or psychiatric disorders. All participants gave written informed consent and the study was approved by ethical board of Ain Shams University.

The Unified Parkinson’s Disease Rating Scale (UPDRS) was used to measure the severity of PD (Lang and Fahn, 1989). The UPDRS for all patients, including PD-OFF, was measured prior to the testing session when all PD patients were on medication. There was no difference in UPDRS between the three patient groups ($F(2,53)=0.29, p=0.75$).

The PD-OFF group was withdrawn from medications for a period of at least 18 hours. The majority of on medication patients were taking dopamine precursors (levodopa-containing medications) and D2 receptor agonists. Specifically, all participants in the PD-ON-ICD group and 14 participants in the PD-ON group were taking D2 agonist medications (either Requip® or Mirapex®). In addition to D2
agonist medications, 10 patients in the PD-ON-ICD group and 11 patients in the
PD-ON group were taking levodopa medications.

All participants were screened for intact cognitive function and absence of dementia
with the Mini-Mental Status Exam (MMSE) (Folstein et al., 1975). Participants
required a score of at least 26 to be considered for the study. All groups were
matched for age and education. In addition, we found no difference between the
groups on the: North American Adult Reading Test (Uttl, 2002), Beck Depression
Inventory (Beck et al., 1987), Mini-Mental Status Exam, forward and backward
digit span tasks (all p-values >0.05, one-way ANOVA). All scales were administered
by trained experts (Table 1).

Table 1  Demographic data. Means are shown with standard deviations in
parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Healthy</th>
<th>PD-OFF</th>
<th>PD-ON</th>
<th>PD-ON-ICD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>66.45(4.70)</td>
<td>63.92(3.99)</td>
<td>63.33(3.98)</td>
<td>64.38(3.32)</td>
</tr>
<tr>
<td>Disease Duration</td>
<td>NA</td>
<td>9.72(2.64)</td>
<td>8.87(3.14)</td>
<td>9.63(2.45)</td>
</tr>
<tr>
<td>HYS</td>
<td>NA</td>
<td>2.54(0.61)</td>
<td>2.40(0.57)</td>
<td>2.47(0.50)</td>
</tr>
<tr>
<td>UPDRS</td>
<td>NA</td>
<td>20.36(5.49)</td>
<td>19.60(6.42)</td>
<td>19.00(5.32)</td>
</tr>
<tr>
<td>NAART</td>
<td>36.25(9.15)</td>
<td>34.64(10.80)</td>
<td>35.60(12.93)</td>
<td>38.00 (6.32)</td>
</tr>
<tr>
<td>MMSE</td>
<td>27.65(1.18)</td>
<td>27.48(0.96)</td>
<td>27.00(0.93)</td>
<td>27.19(1.11)</td>
</tr>
<tr>
<td>Forward DS</td>
<td>6.25(1.65)</td>
<td>6.80(1.66)</td>
<td>6.53(2.13)</td>
<td>6.75(1.69)</td>
</tr>
<tr>
<td>Backward DS</td>
<td>6.25(1.59)</td>
<td>6.32(1.80)</td>
<td>6.47(2.17)</td>
<td>7.00(1.37)</td>
</tr>
<tr>
<td>BDI</td>
<td>7.75(1.97)</td>
<td>6.92(1.32)</td>
<td>8.00(1.69)</td>
<td>6.75(1.57)</td>
</tr>
<tr>
<td>BIS*</td>
<td>54.15(4.51)</td>
<td>56.80(4.74)</td>
<td>57.67(4.18)</td>
<td>61.88(4.56)</td>
</tr>
</tbody>
</table>

Abbreviations: HYS: Hoehn–Yahr scale; NA: not applicable; UPDRS: Unified Parkinson’s Disease Rating
Scale; NAART: North American Adult Reading Test; MMSE: Mini-Mental Status Exam; DS: Digit Span; BDI:
Beck Depression Inventory; BIS: Barratt impulsiveness scale. Asterisks indicate p<0.001.

The diagnosis of ICD was assessed with interviews conducted by neurologists at
Ain Shams University Hospital and associated clinics. ICDs reported included
compulsive shopping (10 patients), hypersexuality (9 patients), gambling (6
patients) and binge eating (4 patients). The majority of participants had more than
one type of ICD (4 patients with only one type of ICD, 11 patients with two ICDs
and 1 patient with 3 ICDs). The Barratt Impulsiveness Scale (BIS) was administered
to measure trait impulsivity in all groups. There was a highly significant difference
in BIS scores between the groups (F(3,72)=8.76, p<0.001). A post-hoc t-test revealed
that the effect was mainly driven by a higher impulsivity in the PD-ON-ICD group. BIS scores for this group were significantly higher than those for the other three groups (\(p<0.02\) for all three tests, two-tailed t-test). We also found a significantly higher BIS scores in the PD-ON group compared with the healthy group (\(p<0.05\), two-tailed t-test).

**Task**

All participants were administered a probabilistic reward and punishment learning task (Figure 1A) (Bódi et al., 2009). On each trial, participants viewed one of four different stimuli (S1, S2, S3 and S4), and were asked to decide whether the stimulus belonged to category A or B. Two stimuli (S1 and S2) were used in the reward-learning trials (win or no-win) and the other two stimuli (S3 and S4) were used in the punishment-learning trials (lose or no-lose). Participants received an outcome after making their choices. There was an optimal choice for each stimulus, which predominately resulted in obtaining reward or avoiding punishment (positive feedback, Figure 1B). Thus, in reward trials, an optimal choice resulted in +25 points 80% of the time and resulted in no reward for 20% of trials. In contrast, a nonoptimal response resulted in +25 points 20% of the time and otherwise resulted in no reward. In punishment trials, an optimal response resulted in -25 points with 20% probability and otherwise resulted in no punishment. In contrast, a nonoptimal response resulted in -25 points 80% of the time and otherwise resulted in no punishment. The task had 160 trials and the order in which stimuli were presented was pseudo-randomized in blocks of 40 trials. For every block, each stimulus was randomly presented in 10 trials.

**Theoretical framework**

We used computational modeling to investigate the mechanistic differences in learning between participant groups. We fitted different RL models to each participant’s choice data. These models were variants of either the Q-learning or the actor-critic framework. Notably, Q-learning and actor-critic frameworks employ different strategies to calculate the PE, the pivotal signal in learning within both frameworks. While the Q-learning framework computes the PE signal based on the estimated value of stimulus-action pairs, the actor-critic framework computes the PE based on the estimated value of stimuli, regardless of the action taken. The different claims of PE computations in these two frameworks can be examined in a relatively theory-neutral manner through model-independent estimation of PE. We also fitted different models to participants’ choices and compared them using Bayesian model comparison. All models use the sequence of choices and feedbacks for every participant in order to estimate the probability of action taken on every trial.
Reinforcement learning models

The first model is the Q-learning model with different learning rates for positive and negative PEs (dual-$\alpha$ QL, Figure 2A). This model learns the value associated with each stimulus-action pair, $Q_t(s_t, a_t)$, using a PE signal, which is the discrepancy between the outcome (reward or punishment) and $Q_t(s_t, a_t)$:

$$\delta_t = o_t - Q_t(s_t, a_t)$$

where $o_t$ is the outcome on trial $t$. The model then updates the current estimated value with the PE:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha^+ \delta_t \text{ if } \delta_t > 0$$
$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha^- \delta_t \text{ if } \delta_t > 0$$

where $\alpha^+$ and $\alpha^-$ are the learning rates for positive and negative PEs, respectively. These learning rates determine the degree that recent PEs affect the estimated value. If $\alpha^+ > \alpha^-$, the effect of positive PEs on learned values is larger than that of

---

**Figure 1** Probabilistic learning task. A) On each trial, one of four stimuli is presented and participant is asked to choose whether the stimulus belongs to category A or B to avoid punishment or obtain reward. B) Structure of the task. For each stimulus, one of the choices (the optimal action) predominantly (80% of the time) results in a positive feedback (either through obtaining a reward or avoiding punishment). The other choice (the nonoptimal action) predominantly results in a negative feedback.
negative PEs, and vice versa if $\alpha^+ < \alpha^-$. The effect of positive and negative PE is equal for $\alpha^+ = \alpha^-$. Frank et al. (2004) hypothesized that different types of dopamine receptors within the striatum mediate the ability to learn from positive and negative PEs via modulation of dopamine activity in the direct and indirect cortico-striato-thalamic pathways, respectively. According to Frank et al. (2004), the positive PE increases phasic dopamine release resulting in learning through D1 receptors. The negative PE, on the other hand, causes a dopamine dip below baseline resulting in learning through D2 receptors (also see Moustafa et al. (2013)).

The probability of choosing each action is computed using the soft-max equation:

$$p(c_t, A | s_t) = \frac{1}{1 + \exp[-\beta Q_t(s_t, A) - Q_t(s_t, B) - \phi(C_t(s_t, A) - (C_t(s_t, B))]}$$

$$p(c_t, B | s_t) = 1 - p(c_t, A | s_t)$$

where $p(c_t = A | s_t)$ and $p(c_t = B | s_t)$ are the probability of choosing A and B, respectively. $\beta$ is the inverse-temperature parameter, which encodes decision noise. $C_t(s_t, A)$ and $C_t(s_t, B)$ represents the choice of A and B on the last presentation of $S_t$, respectively (Lau and Glimcher, 2005; Rutledge et al., 2009). Thus, $C_t(s_t, A) = 1$ and $C_t(s_t, B) = 0$ if A has been chosen in the previous presentation of $S_t$ before trial $t$, but if B has been chosen, $C_t(s_t, A) = 0$ and $C_t(s_t, B) = 1$. Therefore, $\phi$ determines the extent to which the previous choice, independent of reward history, affects the current choice. While positive values of $\phi$ represent a tendency to perseverate on previous choices, negative values represent a tendency to switch more frequently between available options.

The second model is the actor-critic model (standard AC, Figure 2B), which assigns learning and action selection to two different modules. The PE signal in this model is computed based on stimulus values, regardless of the action taken:

$$\delta_t = a_t - V_t(s_t)$$

where $V_t(s_t)$ is the current critic’s value for $s_t$. The critic’s value is then updated using the PE:

$$V_{t+1}(s_t) = V_t(s_t) + a_c \delta_t$$

where $a_c$ is the critic’s learning rate. The PE is also conveyed to the actor in order to update the action value of the selected action in the actor:
Learning in PD Patients with Impulse Control Disorders

\[ Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + a_a \delta_t \]

where \( a_a \) is the actor’s learning rate. Here, if \( \alpha_c > \alpha_a \) then the effect of PEs on the critic is larger than that of actor, and vice versa if \( \alpha_c < \alpha_a \). Note that this is common practice in machine learning that the update of the actor is slower than that of the critic to ensure that the critic has sufficient time to evaluate the current policy (Grondman et al., 2012). However, we enforce no constraints on the critic’s and actor’s learning rates. If participants employed an actor-critic strategy, we would expect that the fitted parameters satisfy this condition for the majority of participants. The probability of each action is computed according to the actor’s action values. A similar soft-max equation as the previous model, dual-\( \alpha \) QL, is used to generate the probability of actions based on actor’s action values and choice perseveration.

The third model is the dual-\( \alpha \) AC model, which is very similar to the standard AC model (Figure 2C). The difference between these two models is how they update stimulus and action values. The dual-\( \alpha \) AC model updates stimulus values through two different learning rates, one for positive PEs and one for negative PEs:

\[ V_{t+1}(s_t) = V_t(s_t) + \alpha_c^+ \delta_t \text{ if } \delta_t > 0 \]
\[ V_{t+1}(s_t) = V_t(s_t) + \alpha_c^- \delta_t \text{ if } \delta_t < 0 \]

If \( \alpha_c^+ > \alpha_c^- \), then the effect of positive PEs on the stimulus value is larger than that of negative PEs, and vice versa if \( \alpha_c^+ < \alpha_c^- \). The actor’s action value is also updated through the two different learning rates for positive and negative PEs:

\[ Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + a_a^+ \delta_t \text{ if } \delta_t > 0 \]
\[ Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + a_a^- \delta_t \text{ if } \delta_t < 0 \]

Here, if \( a_a^+ > a_a^- \) the effect of positive PEs on the actor’s action value is larger than that of negative PEs, and vice versa if \( a_a^+ < a_a^- \). The values for all models were initiated at zero.

Model-independent estimation of prediction error

In this section, we derive a model-independent estimator of PE. This estimator could then be used to assess learning strategies employed by participants in a theory-neutral manner.
Figure 2 The diagram of the three reinforcement learning models. The environment provides three signals for each model: $s$, indicating the current stimulus, $A$, indicating the set of available actions, and $o$, indicating the outcome after receiving the selected action, $a$, from the model. Every model learns appropriate actions by computing a PE signal (indicated by PE block in the diagram) and selects appropriate actions using estimated Q-values of the available set of actions, $A$. A) The dual-α Q-learning model: this model calculates PEs based on the estimated value of stimulus-selected action pair, $Q(s,a)$. Q-values are updated through two different learning rates, $\alpha^+$ and $\alpha^-$, for positive and negative PEs, respectively. B) The standard actor-critic framework: the critic calculates the PE, $\delta$, based on the stimulus value, $V(s)$, independently from the selected action, $a$. The actor computes action values, $Q$, and selects appropriate action, $a$, from a set of available action, $A$, using actor’s Q-values. Both stimulus and action values are updated using the same PE. C) The dual-α actor-critic model: this model has critical features of the previous models. Similar to the standard actor-critic model, the PE is computed based on stimulus values, $V(s)$, independently from the action, $a$, selected by the actor. Similar to the dual-α Q-learning model, this model updates both the critic’s stimulus values, $V$, and the actor’s action values, $Q$, through two different learning rates for positive and negative PEs in the critic, $\alpha^+_c$ and $\alpha^-_c$ and in the actor, $\alpha^+_a$ and $\alpha^-_a$. 
RL models often assume that choices are generated using a soft-max equation of action values:

\[
p_t(a) = \frac{\exp(\beta Q_t(a))}{\exp(\beta Q_t(a)) + \exp(\beta Q_t(a'))}
\]

where \( a \) and \( a' \) are two available choices and \( \beta \) is the inverse-temperature parameter. \( Q_t(a) \) is the action value for \( a \) on trial \( t \), which could be generated by either an actor-critic model or by a Q-learning model. Note that \( Q_t \) is also a function of state (stimulus) in all the models. For simplicity (without loss of generality), we focus on sequence of choices related to one state and omit state in the notation in this section. The probability of taking action \( a' \) on trial \( t \) is computed using a similar equation. Therefore:

\[
\frac{p_t(a)}{p_t(a')} = \frac{\exp(\beta Q_t(a))}{\exp(\beta Q_t(a'))}
\]

Equation 1

Without loss of generality, we suppose that \( a \) is taken at \( t \). Then, the action value of \( a \) should be updated using the PE, \( \delta_t \):

\[
Q_{t+1}(a) = Q_t(a) + \alpha \delta_t
\]

where \( \alpha \) is the learning rate. There is no change in the action value of the other action: \( Q_{t+1}(a') = Q_t(a') \). Therefore:

\[
\frac{p_{t+1}(a)}{p_{t+1}(a')} = \frac{\exp(\beta Q_t(a) + \beta a \delta_t)}{\exp(\beta Q_t(a'))}
\]

Equation 2

By subtracting the logarithm of Equation 1 from the logarithm of Equation 2, we obtain:

\[
\log \frac{p_{t+1}(a)}{p_{t+1}(a')} - \log \frac{p_t(a)}{p_t(a')} = \beta a \delta_t
\]

Equation 3

We define \( n_t(a) \) as the number of times that \( a \) has been chosen in trials \( t' \leq t \). Similarly, \( n_t(a') \) is defined as the number of times that \( a' \) has been chosen in trials \( t' \leq t \). The probability of each choice can be estimated using these variables:

\[
p_t(a) \approx \frac{n_t(a)}{n_t(a) + n_t(a')}
\]
Accordingly, if \( n_t(a') \neq 0 \), Equation 3 can be estimated as:

\[
\alpha \beta \delta_t \approx \delta_t = \log \frac{n_{t+1}(a)}{n_{t+1}(a')} - \log \frac{n_t(a)}{n_t(a')}
\]

where \( a \) is the action taken at \( t \) and \( \epsilon_t \) is the estimator of the PE, which is a quantity that is independent of any specific learning strategy and is purely based on the sequence of choices. Note that the predictions of this estimator match well with the concept of PE. First, if \( a \) is chosen in trials \( t \) and \( t+1 \), \( \epsilon_t \) is positive, suggesting that choosing \( a \) resulted in a positive feedback and increased the probability of choosing \( a \) for subsequent trials. If \( a \) is chosen at \( t \), but not at \( t+1 \), \( \epsilon_t \) is negative, suggesting that choosing \( a \) resulted in a negative feedback and a reduced the probability of choosing \( a \) for future trials. Also, the magnitude of \( \epsilon_t \) is smaller for larger amounts of \( n_t(a) \), which is consistent with the idea that the magnitude of PEs should decrease over time.

**Subjective utility and non-learning models**

We also fitted four additional models to participants' choices in order to investigate whether nonlinearity in subjective values of different outcomes, or some non-learning strategies, could explain data better than the previously mentioned RL models.

**Utility models.** We considered two utility models. These models test the hypothesis that participants' choices can be explained by nonlinearity in subjective value of outcomes. For the probabilistic learning task used in our study, the subjective value refers to the different subjective utilities for reward and punishment.

The first model is the utility QL model as implemented by Niv et al (2012). In this model, the PE is computed based on a nonlinear function of the outcomes:

\[
\delta_t = U(s_t) - Q_t(s_t, a_t)
\]

where \( U(s) \) is the subjective utility of outcome at time \( t \). The action value is then updated using this PE:

\[
Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha \delta_t
\]

As similar to Niv et al. (2012), to model the subjective utility of the outcome, we assumed (without loss of generality) that \( U(0)=0 \), \( U(-25)=-25 \) and \( U(+25)=25u \), where \( u \) is a free parameter that determines the subjective utility of outcome. Values of \( u \) that are smaller than 1 are consistent with hypersensitivity to punishment,
whereas values of $u$ that are larger than 1 are consistent with hypersensitivity to reward. Note that reward and punishment are different from positive and negative PEs that could occur in both reward and punishment trials. This model computes the probability of each action in the same way as the dual-$\alpha$ QL model.

It is also possible to define a subjective utility version of the actor-critic model (utility AC). In this model, the PE is:

$$\delta_t = U(o_t) - V_t(s_t)$$

This PE would then be used by the critic and the actor to update stimulus and action values, respectively. Again, we assumed that $U(0)=0$, $U(25)=-25$ and $U(+25)=25u$, where $u$ is a free parameter that determines the subjective utility of outcome. Similar to the standard AC, two different learning rates are used to update the critic’s stimulus values and the actor’s action values. This model computes the probability of each action in the same way as the standard AC.

**Win-stay lose-shift model.** We also considered a model that implemented a win-stay, lose-shift (WSLS) strategy. This model selects actions based only on the most recent outcome. The WSLS strategy selects the same action that led to success on the next trial or chooses a different option on the next trial when an action did not lead to a success. This strategy can be stochastically modeled using a sigmoid function:

$$p(a_t, a | s_t) = \frac{1}{1 + \exp(-\beta w_t(s))}$$

where $a$ is the chosen action in the previous presentation of $s_t$ and $\beta > 0$ encodes decision noise. To model the WSLS strategy, we assumed (without loss of generality) that $w_t = -1$ if the previous presentation of $s_t$ was a lose trial and $w_t = W$ if it was a win trial. $W>0$ is the parameter that determines the weight of win compared to loss. If $W>1$, the effect of win on the subsequent choice is larger than that of loss, and vice versa if $W<1$. The effect of win and loss on subsequent choices is symmetric if $W=1$. For all positive values of $W$, the probability of choosing the same action as the previous trial is more than the alternative action if the previous trial was a win trial and less than the alternative action if the previous trial was a loss trial. Note that in our probabilistic learning task, win trials were those that resulted in obtaining a reward in reward trials or avoiding a punishment in punishment trials. We fitted two WSLS models to participants’ choices. For the first model we assumed both $\beta$ and $W$ were free, and in the second one we fixed $W$ at 1. The values for all models were initiated at zero.
Model fitting procedure
We used a hierarchical Bayesian procedure for fitting models to participants’ choices as described in Huys et al. (2011a, 2012). All parameters of the models are assumed to be free (see Table 2 for the number of free parameters in each model) except for $\beta$ in the three actor-critic models (standard AC, dual-$\alpha$ AC and utility AC), which was fixed at 1. This is because the probabilities of choices for these models are affected by the product of the learning rate parameter of the actor and $\beta$ and this is the only way that these parameters affect the likelihood function. These two variables are indeed co-linear. To show that fixing $\beta$ at 1 is statistically justified, we also fitted these models with $\beta$ as a free parameter and used the likelihood ratio test to examine whether these models fit significantly better than the same models with $\beta$ fixed at 1. For all three models, the fits were not significantly improved by having $\beta$ as a free parameter ($p>0.9$ for all groups, likelihood ratio test). Accordingly, the standard AC, dual-$\alpha$ AC and utility AC models have 3, 5 and 4 free parameters, respectively.

In the hierarchical Bayesian procedure, the parameters of an a priori distribution for individual parameters were estimated using participants’ choices through the Expectation Maximization (EM) algorithm (Dempster et al., 1977). This algorithm is a well-known method for finding maximum a posteriori, which alternates between an expectation step and a maximization step. We used Laplace approximation (MacKay, 2003) for the expectation step on each iteration. Assuming a normal distribution for individual parameters, $\Theta_i$ for $i$th participant, this method estimates the mean and the variance of the distributions across the whole group, $\Theta$, which serves as an a priori distribution for finding the maximum a posteriori on the next iteration. For example, for the dual-$\alpha$ AC model, the group parameters are:

$$\Theta = [\mu_\alpha, v_\alpha, \mu_\alpha^*, v_\alpha^*, \mu_\alpha^{-*}, v_\alpha^{-*}, \mu_\alpha^{-*}, v_\alpha^{-*}, \mu_\phi, v_\phi]^T$$

where $\mu$ and $v$ indicate the mean and deviance of the corresponding parameter, respectively. The group mean and variance were estimated separately for each group and were used to define an a priori Gaussian distribution for individual parameters. Thus, four sets of parameters, associated with four groups, were estimated. For the details of the hierarchical fitting procedure please refer to Huys et al. (2012).

Bayesian model selection
We employed a Bayesian model selection approach to assess which model better captures participants’ choices. This approach selects the most parsimonious model.
by balancing between model fits and different levels of complexity of the models (Kass and Raftery, 1995; MacKay, 2003).

We computed approximate model evidence, \( P(D | M) \), which is the probability of participants’ choices, \( D \), given the model \( M \). We approximated \( P(D | M) \) (in log-space) using Bayesian Information Criterion (BIC):

\[
- \log P(D | M) \approx - \log P(D | M, \Theta_{ML}) + \frac{1}{2} |\Theta| \log |D|
\]

where \( D \) is the set of all participants’ choices in the group, \( |D| \) is the number of choices for the whole group and \( |\Theta| \) is the number of group parameters. \( \Theta_{ML} \) is obtained using maximum likelihood:

\[
\Theta_{ML} = \arg \max_{\Theta} P(D | M, \Theta)
\]

Since \( \Theta_{ML} \) determines an a priori distribution for individual parameters, we can obtain \( P(D | M, \Theta_{ML}) \) using the Laplace approximation:

\[
- \log P(D | M, \Theta_{ML}) \approx \sum_{i} \log P(D_i | M, \Theta_{ML}, \Theta_{i,MAP}) - \sum_{i} \log P(\Theta_{i,MAP} | \Theta_{ML}) - \frac{1}{2} \sum_{i} |\Theta_i| \log 2\pi + \frac{1}{2} \sum_{i} \log |H_i|
\]

where \( D_i \) is the set of \( i \)th subject’s choices, \( |\Theta_i| \) is the number of free parameters in the model for \( i \)th subject, \( |H_i| \) is the determinant of the Hessian matrix for \( i \)th subject at \( \Theta_{i,MAP} \), and \( \Theta_{i,MAP} \) is the maximum a posteriori of parameters for the \( i \)th subject:

\[
\Theta_{i,MAP} = \arg \max_{\theta} P(D_i | M, \Theta_{ML}, \theta) P(\theta | \Theta_{ML})
\]

**Model selection using cross-validation**

We also performed a cross-validation analysis as a control analysis for model selection. Parameters of the models were fitted based on a subset of choices and generalization of models were assessed by quantifying the prediction probability of the models on a different subset of choices that was not used for fitting (see Daw (2011) for shortcomings of this method in learning studies). Similar to Camerer and Ho (1999), the parameters of models were estimated based on the first two-thirds of trials using the hierarchical Bayesian fitting procedure. Next, the negative log-likelihood of the prediction probability of choices on the remaining one-third of trials was computed and reported.
Statistical analyses
Due to non-Gaussian statistics (since some parameters are expected to lie in the unit range), we used the nonparametric Wilcoxon test for parameter comparison between groups. To ensure that between-group differences were not dependent on parameter regularization used in the hierarchical Bayesian procedure (Wunderlich et al., 2012b), we employed a permutation test approach as a control analysis. For each significant between-group difference, the labels of the groups were randomly permuted 200 times across the participants of both groups. The parameters for these two pseudo random groups were then found using the hierarchical Bayesian procedure. We then tested whether the effect-size in the real data (assessed by the difference in the median of two groups’ parameters) was more than the effect-size for the pseudo random groups.

We also examined between-group differences in stimulus values for both reward and punishment trials. Each subject’s fitted parameter values were used to estimate the value of stimuli. The nonparametric Wilcoxon test was used to test between-group differences. A similar control analysis was also conducted to ensure that the results were not dependent on parameter regularization. Since it is not possible to test between-group differences in stimulus values using the permutation test (due to the dependency of stimulus values in the last presentation of each stimulus on both fitted parameters and sequence of outcomes received), we re-fitted the dual-α AC model to participants’ choices using the hierarchical Bayesian procedure but with only one a priori distribution defined across all participants. Since individual parameters were obtained using the same a priori, the between-group differences cannot be attributed to parameter regularization.

Results
Behavioral data
The probability of optimal responses made by participants was analyzed using an ANOVA with group (4 levels: PD-OFF, PD-ON, PD-ON-ICD and healthy controls) as a between-subject factor and valence (reward or punishment) as a within-subject factor (Figure 3). This analysis revealed a highly significant interaction between group and valence ($F(3.0,72.0)=15.81$, $p<0.001$), as well as a significant main effect of group ($F(3.0,72.0)=3.79$, $p<0.05$), but no significant main effects of valence ($F(1.0,72.0)=2.23$, $p=0.14$). Further analyses with the additional factor block (2 levels: the first half and the second half of the 160 trials) were conducted to assess learning effects. This analysis revealed a significant main effect of block ($F(1.0,72.0)=14.25$, $p<0.001$), but no interaction between block and other factors (refer to Figure 4 for learning curve).
Figure 3 Performance of the four groups on the probabilistic learning task. A) Mean performance in reward trials and B) mean performance in punishment trials. For reward trials, the PD-ON group performed better than the PD-OFF group, but worse than the PD-ON-ICD group. The opposite pattern of performance was observed in punishment trials. Error-bars reflect standard error.

Figure 4 Learning curve for A) reward and B) punishment trials. The 160 trials are divided in 4 blocks. Each block contains 20 reward and 20 punishment trials. Error-bars reflect standard error.
Next we broke down the significant group by valence interaction into simple main effects of group for the reward and punishment trials separately. All p-values are from two-tailed t-test. Thus, reward learning was impaired in the PD-OFF group relative to the other three groups (healthy controls, PD-ON and PD-ON-ICD groups: \( p < 0.001 \), \( p < 0.01 \) and \( p < 0.001 \), respectively). Conversely, the PD-ON-ICD group showed better reward learning than the other three groups (with healthy controls: \( p = 0.015 \); with PD-ON: \( p = 0.016 \)).

The opposite pattern of performance was observed for punishment learning. The PD-OFF group exhibited better punishment learning than the PD-ON-ICD (\( p = 0.003 \)) and PD-ON groups (\( p = 0.046 \)), although there was no significant difference in punishment learning between PD-OFF and healthy participants (\( p = 0.43 \)). Moreover, punishment learning was impaired in the PD-ON-ICD group relative to the healthy control group (\( p = 0.028 \)), although not relative to the PD-ON group (\( p = 0.41 \)).

**Model-independent evaluation of learning strategy**

Two different strategies could be used to compute the learning signal in the probabilistic learning task. First, the PE could be computed based on the outcome received, regardless of which action was taken. This strategy is used by the actor-critic framework. The second strategy is to compute the PE based on the value of the action taken. This strategy is used by the Q-learning framework. The probabilistic learning task allowed us to distinguish between these two learning strategies. For example, if the percentage of optimal responses is 70%, the critic’s stimulus value is affected by the outcomes of both actions; and its value (after sufficient trials) is in the middle of two actions’ values estimated by the Q-learning framework. For a rewarding stimulus such as S1, the Q-learning value of action A (optimal action), the Q-learning value of action B and actor-critic stimulus values are around 20, 5 and 15.5, respectively. Thus, if taking an action results in a positive feedback (an outcome of 25 points), the PE computed by actor-critic is 9.5, but the PE by Q-learning is either 5 or 20, depending on which action is taken. Also, if taking an action results in a negative feedback (an outcome of 0 points), the PE computed by the actor-critic is -15.5, but the PE computed by Q-learning is either -20 or -5, depending on the action selected. Therefore, two key events may influence learning signal in this task: whether feedback was positive or negative, and whether the action taken was optimal or nonoptimal.

Figure 5A and Figure 5B illustrate the simulated learning signal predicted by the Q-learning and actor-critic frameworks, respectively. As these figures show, while both strategies predict a main effect of the feedback, the predictions of the two
frameworks are different in terms of the action. While the actor-critic framework predicts no main effect of action, the Q-learning framework predicts the opposite.

To assess learning strategies employed by participants in a relatively theory-neutral manner, we directly assessed the effects of feedback and action on the model-independent estimated PEs across participants (see Methods), a quantity that is

Figure 5  Factorial analysis of model-independent estimates of the learning signal. A) Q-learning framework computes the learning signal based on action values and predicts that this signal depends on whether optimal action or nonoptimal action is taken; B) Actor-critic framework computes the learning signal based on the stimulus value, regardless of which action is taken. C) Model-independent estimated learning signal based on the data, averaged across participants, is consistent with the prediction of the actor-critic framework. Both models were simulated with learning rates, $\alpha$, of 0.05 and $\beta$ inverse-temperature of 0.1. The learning signal for both models, $\epsilon$, was defined as $\beta \alpha \delta_t$, where $\delta_t$ is the PE computed by the model at trial $t$. See Methods for on the definition of the model-independent estimates of learning signal. Error-bars reflect standard error.
purely based on the sequence of choices for each stimulus. We analysed the model-independent estimated PEs using an ANOVA with feedback and action as within-subject factors and with group as a between-subject factor. This analysis revealed a highly significant main effect of feedback ($F(1.0,70.0)=38.5$, $p<0.001$), consistent with the prediction of both Q-learning and actor-critic frameworks. However, there was no main effect of action ($F(1.0,70.0)=0.37$, $p=0.55$), suggesting that the learning strategy employed by participants is consistent with the actor-critic learning strategy, but not with that of the Q-learning. As predicted by both learning strategies, no interaction between feedback and action was observed ($F(1.0,70.0)=1.29$, $p=0.26$). Also, no main effect of group, as well as no two- or three-way interactions between group and the other factors were observed ($p>0.5$), suggesting that all groups employed the same learning strategy. Therefore, we plotted model-independent estimated learning signal across participants in all groups in Figure 5C.

We further studied the effects of feedback and action separately for each group using an ANOVA with feedback and action as within-subject factors. Consistent with the previous analysis, there was a main effect of feedback in all four groups (all $p$-values $<0.02$). No main effect of action and no interaction were observed for any of the groups (all $p$-values $>0.16$). Taken together, these findings suggest that the learning strategy in all groups is consistent with the predictions by the actor-critic framework.

Note that this analysis holds for the different variants of Q-learning and actor-critic frameworks. Specifically, while the dual-$\alpha$ AC model predicts no main effect of action on the learning signal, the dual-$\alpha$ QL model predicts a main effect of action. Also, both models predict a main effect of feedback and neither predicts an interaction between these factors. Therefore, the results of the analysis of model-independent estimated PEs are consistent with dual-$\alpha$ AC claims about PEs, but not with those of the dual-$\alpha$ QL model.

**Model comparison**

Motivated by these results, we examined the full fit of the models to participants’ choices. First, we verified that the models fit significantly better than chance; they did so at $p<0.001$ for all four groups (likelihood ratio tests). Then, Bayesian model comparison was conducted to identify the best model in each group (Table 2). As Table 2 shows, the negative log-model evidence is lower (with log-Bayes factor of at least 9.5) for the dual-$\alpha$ AC than for the other models for all groups, providing compelling support that the dual-$\alpha$ AC model best captures participants’ choices. In the Bayesian model comparison literature, a log-Bayes factor of more than three
Learning in PD Patients with Impulse Control Disorders

is taken as strong evidence (cf. the \( p<0.05 \) criterion often employed in classical statistics) (Kass and Raftery, 1995; Daw, 2011). As Table 2 shows, the smallest difference in log-evidence between the best (dual-\( \alpha \) AC) and the second best model (dual-\( \alpha \) QL) is the one for the PD-ON-ICD group. Because this group is the critical group in this study, we also employed a cross-validation approach as a control analysis to compare the plausibility of these two models for this group. Thus, parameters were fitted based on the first two-thirds of trials and performance of the models quantified on the remaining unseen one-third of trials (Camerer and Ho, 1999). The negative log-likelihood for the dual-\( \alpha \) AC and the dual-\( \alpha \) QL on the testing dataset were 478.4 and 536.6, respectively. Therefore, the results of cross-validation model selection are consistent with those of the Bayesian model selection, demonstrating strong evidence in favor of the dual-\( \alpha \) AC model.

Subsequently, we simulated choices by the best model, the dual-\( \alpha \) AC model, using the fitted parameters in order to verify that the dual-\( \alpha \) AC model simulates a similar pattern of between-group differences in optimal responses as observed in the behavioral data (plotted in Figure 3). These simulated choices were then subject to the same two-tailed \( t \)-test comparisons employed in the analyses of between-group differences in behavioral performance. Overall, this simulation analysis replicated similar between-group differences as those observed in the empirical data. The performance of the PD-ON-ICD group in reward trials was significantly better than the other groups (\( p<0.01 \)). In punishment trials, the PD-OFF group performed significantly better than the PD-ON-ICD group (\( p=0.025 \), but not when compared

Table 2 Bayesian model selection.

<table>
<thead>
<tr>
<th></th>
<th>No. parameters</th>
<th>Healthy</th>
<th>PD-OFF</th>
<th>PD-ON</th>
<th>PD-ON-ICD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard AC</td>
<td>3</td>
<td>1653.7</td>
<td>2195.8</td>
<td>1217.5</td>
<td>1344.3</td>
</tr>
<tr>
<td>Dual-( \alpha ) QL</td>
<td>4</td>
<td>1660.9</td>
<td>2208.9</td>
<td>1212.8</td>
<td>1299.2</td>
</tr>
<tr>
<td>Dual-( \alpha ) AC</td>
<td>5</td>
<td>1587.9</td>
<td>2091.6</td>
<td>1171.4</td>
<td>1289.7</td>
</tr>
<tr>
<td>Utility QL</td>
<td>4</td>
<td>1687.5</td>
<td>2182.9</td>
<td>1239.8</td>
<td>1306.7</td>
</tr>
<tr>
<td>Utility AC</td>
<td>4</td>
<td>1657.4</td>
<td>2180.1</td>
<td>1212.8</td>
<td>1303.6</td>
</tr>
<tr>
<td>WSLS</td>
<td>2</td>
<td>1893.8</td>
<td>2499.9</td>
<td>1414.3</td>
<td>1531.1</td>
</tr>
<tr>
<td>WSLS (fixed ( W ))</td>
<td>1</td>
<td>2124.0</td>
<td>2742.6</td>
<td>1587.1</td>
<td>1683.5</td>
</tr>
</tbody>
</table>

These numbers represent the negative log-likelihood of data in the corresponding group given the associated model. The Bayesian model selection takes into account both the goodness of fit and the generalizability of the models. Lower values are associated with better fits. The dual-\( \alpha \) AC model fits better than other models for all four groups. Abbreviations: Q-learning (QL); actor-critic (AC); win-stay lose-shift (WSLS).
to the other two groups ($p>0.5$). Also, consistent with the behavioral results, no difference was found between the PD-ON-ICD and the PD-OFF groups in punishment trials ($p=0.24$). The simulated choices failed to replicate the findings of significant lower performance by PD-OFF compared with healthy controls and PD-ON in reward trials ($p>0.05$), although the mean performance of PD-OFF was lower than these groups in reward trials.

**Between-group differences in the critic and actor**

Next, we assessed between-group differences in parameter values of the best model, dual-$\alpha$ AC. Figure 6 shows the learning rates in the critic and actor. As this figure shows, the actor’s learning rates are generally lower than the critic’s learning rates. This learning rate profile ensures that the critic has sufficient time to evaluate the current policy exploited by the actor (Grondman et al., 2012).

First, we studied between-group differences in the critic’s parameters. According to our hypothesis, we expected an association between ICD and the critic’s learning rates. Although there was no significant difference in $\alpha_c^+$ between PD-ON-ICD and other groups ($p>0.1$ for all three tests, Figure 6A), we found a significantly lower learning rate from negative PEs in PD-ON-ICD. Indeed, as Figure 6B shows, $\alpha_c^-$ in PD-ON-ICD was less than healthy participants ($p=0.002$), PD-OFF ($p<0.001$) and PD-ON ($p=0.017$). No other group differences in $\alpha_c^-$ were found.

We also investigated between-group differences in the actor’s learning rates. Based on the previous data (Frank et al., 2004) and our hypothesis that PD is associated with action valuation deficits, we expected a relatively lower learning rate for the positive PE in PD-OFF and a relatively lower learning rate for the negative PE in PD-ON. As Figure 6C shows, $\alpha_a^+$ was significantly lower in PD-OFF than PD-ON ($p=0.050$). Conversely, $\alpha_a^-$ was higher in PD-OFF than PD-ON, despite showing only a trend towards significance ($p=0.058$, Figure 6D). Consistent with our hypothesis, there was no significant difference between PD-ON-ICD and PD-ON in terms of the actor’s parameters (no difference between PD-ON and PD-ON-ICD for either $\alpha_a^+$ ($p=0.35$) or $\alpha_a^-$ ($p=0.77$)).

Using the actor-critic framework, it is possible to also evaluate stimulus values. Thus, we derived the value of every stimulus at the end of the task (the last presentation of the stimulus) for each subject using the subject’s choices and the fitted parameters in the dual-$\alpha$ AC model (Figure 7). We then tested between-group differences in stimulus value separately in reward and punishment trials. Note that two stimuli were only presented in reward trials and two other stimuli were only presented in punishment trials. The stimulus value in punishment trials for
the PD-ON-ICD group was significantly less negative than those for the PD-OFF (p=0.003), PD-ON (p=0.038) and healthy control (p=0.02) group, suggesting that PD patients with ICD underestimate the adverse consequences of stimuli associated with punishment. No significant difference in the stimulus value in reward trials between PD-ON-ICD and PD-ON was found (p=0.35). Consistent with our hypothesis, the two groups of PD patients without ICD showed a similar pattern of stimulus values in both reward and punishment trials (no difference between PD-OFF and
Figure 7 Stimulus value in reward and punishment trials. The stimulus values were obtained using the fitted parameters in the dual-$\alpha$ AC model for the last presentation of each stimulus and averaged across participants. PD-ON-ICD patients exhibited significantly less negative stimulus value in punishment trials compared with the other groups. Error-bars reflect standard error.

PD-ON for either reward ($p=0.74$) or punishment ($p=0.60$) trials), which supports the idea that PD is not associated with stimulus valuation deficits.

We should note that our main results are independent of the parameter regularization: the critic’s learning rate for the negative PE, $\alpha_c^-$, was significantly lower in the PD-ON-ICD group than in other three groups even when using the permutation test ($p<0.05$, two-tailed test). The control analysis for between-group differences in stimulus values also revealed the same significant between-group differences as in our original analysis.

Although the dual-$\alpha$ AC model outperformed the dual-$\alpha$ QL model in all four groups, we also present the results of the between-group difference tests in learning rates for the positive and negative PEs in the dual-$\alpha$ QL model to highlight the benefits of AC modeling for ICD. The learning rate for positive PEs, $\alpha^+$, was significantly higher in PD-ON compared to healthy control ($p=0.002$) and marginally higher in PD-ON compared to PD-OFF ($p=0.07$). This parameter was also significantly higher in PD-ON-ICD compared to healthy controls ($p<0.001$). However, there was no difference in $\alpha^+$ between PD-ON-ICD and PD-ON ($p=0.51$). There was also no difference between PD-ON-ICD and PD-OFF ($p=0.15$). No significant between-group differences found in the learning rate for negative PEs, $\alpha$ (all
Therefore, as these analyses revealed, no difference was found in parameter values of the dual-α QL between on medication patients with ICD and those without ICD.

**Between-group differences in the perseveration**

A recent RL study of PD patients reported that the perseveration parameter is dopamine-dependent. Thus, off medication PD patients exhibited higher perseveration than on medication patients (Rutledge et al., 2009). Although it is not the main focus of this study, we also examined the effects of the perseveration parameter on model fits and between-group differences in $\varphi$, the parameter determining the degree that perseveration affects choice. To show that including $\varphi$ in the dual-α AC model is statistically justified, we first tested whether the model with the perseveration parameter fitted significantly better than the same model without the perseveration parameter; it did so for all groups ($p<0.001$, likelihood ratio test). Note that the perseveration parameter encodes the probability of repeating an action on the subsequent presentation of a stimulus. An alternative way to define perseveration could be to compute the probability of repeating an action on the subsequent trial regardless of the stimulus presented. Therefore, we also fitted a dual-α AC model with the stimulus-independent perseveration and used model selection to test whether the original model outperforms this model; it did so for all groups (with log-Bayes factor of more than 6.3).

As Figure 8 shows, consistent with Rutledge et al. (2009), we found significantly higher perseveration values in the PD-OFF group compared with healthy controls ($p=0.03$). Interestingly, we also found significantly lower perseveration in the PD-ON-ICD group than in the PD-OFF, PD-ON and healthy control groups ($p<0.01$ for all three tests). No significant difference was found between the PD-OFF and the PD-ON groups ($p=0.08$).
Dopaminergic medications trigger ICD in a subset of PD patients. In this study, we used a reward and punishment probabilistic learning task and fitted RL models to participants’ choices to investigate the mechanistic differences in stimulus valuation and action selection in PD patients with and without ICD. The probabilistic learning task allowed us to distinguish between different learning strategies used by QL and AC frameworks through their different claims about the effects of actions taken on learning. We found that model-independent estimates of the learning signal are consistent with the hallmark of the AC learning strategy. The full fit of models and Bayesian model comparison revealed that an AC model (with different learning rates for positive and negative PEs in both the critic and the actor) best matches participants’ choices.

We found that PD patients with ICD (on medication) are more sensitive to rewarding outcomes. Computational modeling revealed that these patients also underestimate adverse consequences of stimuli associated with punishment. We also found computational evidence that patients with ICD exhibit reduced ability in updating stimulus values by negative PEs. Therefore, our findings suggest that distorted stimulus valuation could result in aberrant PE signals, which subsequently affects action values.
There is a great deal of evidence that the ventral striatum contributes to decision making in a manner consistent with the role of the critic in stimulus valuation and PE computations (Cardinal et al., 2002; Dayan and Balleine, 2002; Packard and Knowlton, 2002; O’Doherty et al., 2004). Therefore, our findings are in line with previous studies that found dopamine-dependent ventral striatal dysfunction in PD patients with ICD symptoms (Cools et al., 2007a; Dagher and Robbins, 2009; Steeves et al., 2009; Voon et al., 2010). For example, in a [11C] raclopride positron emission tomography study of PD patients with and without pathological gambling, Steeves et al. (2009) found greater decreases in binding potential in the ventral striatum in on medication PD patients with pathological gambling. In addition, Voon et al. (2010) reported impaired PE signaling in the ventral striatum of PD patients with ICD.

We also found that PD (without ICD) is associated with parameters related to action valuation, but not with stimulus valuation. Therefore, although PD patients without ICD exhibited no deficit in learning stimulus value used for calculating PEs, they showed abnormalities in updating action values with the information signaled by the critic. Therefore, our findings suggest that PD patients without ICD have relatively intact PE computations (in their relatively intact ventral striatum), but the effects of PEs on action values are distorted (in their severely depleted dorsal striatum). These findings are consistent with the hypothesis that the dorsal striatum, the most affected striatal region in PD, is responsible for action valuation and selection. In addition, we also found that the action valuation abnormalities in PD patients without ICD interact with dopaminergic medications. Therefore, consistent with previous data (Frank et al., 2004; Moustafa et al., 2008, 2013; Bódi et al., 2009), we found that whereas off-medication PD patients were better at learning from punishment, on-medication PD patients were better at learning from reward. Mechanistically, we found that off-medication patients, compared with on-medication patients, showed lower action value learning from positive PEs and marginally higher action value learning from negative PEs. Notably, almost all patients in this study received D2 agonist medications, which stimulate D2 dopamine receptors. Therefore, this finding is consistent with the hypothesis of Frank et al. (2004) that different types of dopamine receptors within the striatum, especially those in more dorsal regions, mediate the ability to learn from positive and negative PEs via modulation of dopamine activity in the direct and indirect basal ganglia pathways, respectively (Frank et al., 2004, 2007a; O’Reilly et al., 2007). According to this hypothesis, the positive PE increases phasic dopamine release, which facilitates learning by acting on D1 receptors. Conversely, the negative PE results in a dopamine dip below baseline, which facilitates learning by acting on D2 receptors.
Although the role of D1 and D2 receptors in the ventral striatal region, especially the nucleus accumbens shell, is less clear than in the dorsal striatum (Ikemoto et al., 1997; Hopf et al., 2003), there is increasing evidence that the ventral striatal D2 receptors are also involved in learning from negative PEs. Indeed, the negative PE results in dopamine dips below baseline (Bayer and Glimcher, 2005; Hart et al., 2014), which can stimulate high-affinity D2 receptors, but not D1 receptors (Frank et al., 2004). It has also been suggested that D2, but not D1, receptors are stimulated with tonic dopamine release (Grace, 1991). Therefore, as noted by Frank et al. (2004), D2 agonist drugs might fill the dips and reduce the ability to learn from negative PEs. In rats, nucleus accumbens D2 stimulation with a dopamine agonist reduced the ability to learn from negative feedback (Goto and Grace, 2005). In addition, Al-carriers of TAQ-1A polymorphism, which is associated with a lower density of striatal D2 receptors, showed impaired learning from negative feedbacks and aberrant reward-related responses in the ventral striatum (Klein et al., 2007). This hypothesis is consistent with data reporting that ICDs are observed more often in patients on D2 agonist medications (Weintraub et al., 2006; Voon et al., 2007).

An important open question is which individual differences in PD patients with ICD interact with D2 agonist medications and induce compulsive behaviors? One possible answer is that patients vulnerable to ICD have a lower ventral striatal D2 receptor density, even before the onset of PD (Dagher and Robbins, 2009). There is limited but important evidence from animal models of cocaine addiction that rats with lower nucleus accumbens D2 receptor density are more impulsive, even prior to cocaine exposure (Dalley et al., 2007), and are more likely to develop compulsive drug seeking (Belin et al., 2008). In addition, Weintraub et al. (2006) investigated ICD in a large sample of PD patients and reported that those with ICDs were more likely to have had ICDs before the onset of PD. Moreover, animal model studies of addiction have reported that drug exposure further reduces striatal D2 receptors (Nader et al., 2002; Porrino et al., 2004a). Similarly, the overstimulation of the ventral striatum in PD patients by D2 agonist medications may further reduce the density of ventral striatal D2, making them more susceptible to develop ICD. Consistent with these ideas, it has been reported that PD patients with ICD showed lower density of D2 receptors in the ventral striatum (Steeves et al., 2009), although it is not clear from this particular study that the reduced level of D2 receptors in the ventral striatum is a predisposing neurobiological trait and/or a consequence of medication.
Conclusion
In summary, we found that whereas PD is associated with parameters related to action valuation and selection, ICDs in PD are mechanistically associated with parameters related to stimulus valuation and PE computations. Specifically, we found computational evidence that ICDs in PD are associated with lower learning rates from negative feedbacks in the critic. These findings offer a computational interpretation of ICDs in PD and highlight the value of computational modeling in understanding cognitive deficits associated with psychiatric disorders (Redish et al., 2008; Huys et al., 2011b; Maia and Frank, 2011; Montague et al., 2012; Monterosso et al., 2012).
4

CHAPTER

Human choice strategy varies with anatomical projections from ventromedial prefrontal cortex to medial striatum

This chapter is adapted from:
Abstract

Two distinct systems, goal-directed and habitual, support decision making. It has recently been hypothesized that this distinction may arise from two computational mechanisms, model-based and model-free reinforcement learning, neuronally implemented in frontostriatal circuits involved in learning and behavioral control. Here, we test whether the relative strength of anatomical connectivity within frontostriatal circuits accounts for variation in human individuals’ reliance on model-based and model-free control. This hypothesis was tested by combining diffusion tensor imaging with a multistep decision task known to distinguish model-based and model-free control in humans. We found large inter-individual differences in the degree of model-based control, and those differences are predicted by the structural integrity of white-matter tracts from the ventromedial prefrontal cortex to the medial striatum. Furthermore, an analysis based on masking out of bottom-up tracts suggests that this effect is driven by top-down influences from ventromedial prefrontal cortex to medial striatum. Our findings indicate that individuals with stronger afferences from the ventromedial prefrontal cortex to the medial striatum are more likely to rely on a model-based strategy to control their instrumental actions. These findings suggest a mechanism for instrumental action control through which medial striatum determines, at least partly, the relative contribution of model-based and model-free systems during decision-making according to top-down model-based information from the ventromedial prefrontal cortex. These findings have important implications for understanding the neural circuitry that might be susceptible to pathological computational processes in impulsive/compulsive psychiatric disorders.
Introduction

Instrumental actions are controlled by two distinct strategies, a flexible but computationally expensive goal-directed strategy, and a rapid but rigid habitual strategy. This distinction has recently been formalized in a normative computational account in which two reinforcement learning strategies, a “model-based” and a “model-free” system jointly control instrumental actions (Daw et al., 2005). The model-free system directly reinforces actions that lead to reward, ignoring the probabilistic structure of predictive cues in the environment. The model-based system uses an internal model of probabilistic regularities in the environment to evaluate candidate actions.

It is generally assumed that reliance on habitual actions is influenced by state factors. For instance, stress, dual-tasking, administration of dopaminergic drugs, transcranial magnetic stimulation, and striatal presynaptic dopamine affect the relative balance between model-based and model-free control (Wunderlich et al., 2012b; Otto et al., 2013a, 2013b; Deserno et al., 2015). Those state-dependent effects have been indexed by population-level summary parameters, treating inter-individual trait variability as noise. In fact, structural differences in the neural circuits supporting model-free and model-based control might explain inter-individual variability in the relative contribution of those two systems. Accordingly, this study considers whether human choice is systematically biased by stable neuro-anatomical trait factors.

The available evidence suggests that model-based and model-free control systems rely on partly different frontostriatal circuits. Ventromedial prefrontal cortex (vmPFC) and the dorsomedial striatum are implicated in model-based control (Gläscher et al., 2010; Daw et al., 2011; Wunderlich et al., 2012a; Lee et al., 2014), whereas the dorsolateral striatum is implicated in model-free control (Wunderlich et al., 2012a). This neuroanatomical segregation of computational functions nicely overlaps with the long-standing distinction between goal-directed and habitual modes of behavioural control. Works with behaving rodents (Balleine and Dickinson, 1998; Corbit and Balleine, 2003; Yin et al., 2005) and healthy humans (Valentin et al., 2007; Tanaka et al., 2008; Gläscher et al., 2009) have shown that dorsomedial striatum as well as vmPFC are implicated in goal-directed actions. On the other hand, dorsolateral striatum has been shown to contribute to habitual responses (de Wit et al., 2012). Building on this evidence, in this study we tested whether inter-individual variation in the strength of anatomical connectivity within those frontostriatal circuits predicts the relative contribution of model-free and model-based systems to human choice. We hypothesized that inter-subject
variability in the relative balance between model-based and model-free control depends on the integrity of anatomical frontostriatal connections, with the vmPFC and dorsomedial striatum implicated in model-based control and frontal motor areas and dorsolateral striatum implicated in model-free control.

Using probabilistic tractography of diffusion-tensor images (DTI), connectivity-based parcellation of the frontal lobe (Beckmann et al., 2009; Mars et al., 2011; Neubert et al., 2014), and a computationally-explicit learning model of a multistep decision task (Daw et al., 2011), this study mechanistically grounds the balance between model-free and model-based control systems into the relative strength of different frontostriatal loops. To anticipate the results, we found evidence that the structural integrity of white-matter tracts between vmPFC and medial striatum predicts individuals’ reliance on model-based control. By masking out bottom-up tracts, we found evidence that top-down afferences from the vmPFC to the medial striatum determine the relative contribution of model-based control during decision-making.

Methods

Participants
We recruited 33 healthy volunteers. All participants gave informed consent and the study was approved by the local ethics committee. All participants underwent two separate sessions, a diffusion-weighted magnetic resonance imaging scan; and a behavioral session during which subjects were tested on the multistep decision task used previously to quantify model-based and model-free components of instrumental actions in humans (Daw et al., 2011) (Figure 1A). Two participants quit the study after the first session. Thus, data from 31 participants (15 men, 22.7±2.5 mean age) were analyzed. Participants had no history of neurological and psychiatric disorders.

Task
On each trial of the task, subjects first made a choice between two fractal stimuli leading to one of the two different second-stage sets represented by different colors. Participants then made another choice between two stimuli presented in the second-stage set. Each stimulus at the second-stage was associated with a specific probability of delivering a monetary reward. Similar to previous studies with this task (Daw et al., 2011; Smittenaar et al., 2013), the probabilities of delivering reward changed independently and slowly based on a Gaussian random walk to motivate participants to continue learning throughout the task. Critically,
each choice at the first stage led predominantly (70%) to one of the two sets at the second stage (common transition), and less frequently (30%) to the other set (rare transition). This feature of the task allowed us to distinguish contribution of model-based and model-free in choices. The task consisted of 201 trials.

**Behavioural analysis**

Logistic regression was used to analyze responses at the first level of the task independently for each participant. The multistep task has a 2x2 factorial design, where the factors are transition (common or rare) and reward delivery on the previous trial (rewarded or unrewarded). Thus, first-stage choices, encoded as binary stay/switch responses, were regressed against four predictors: main effects of the two factors, interaction effect of the two factors and an intercept representing the tendency to stay with the same choice regardless of transition and reinforcement factors (stickiness). Logistic regression was performed separately for each subject using MATLAB Statistics toolbox (glmfit routine). The degree of model-free and model-based deployment were quantified as the main effect of reward delivery and the interaction effect between reward delivery and transition, respectively.

**Computational modelling**

We also fitted data to reinforcement learning models previously suggested to account for choices in this task (Daw et al., 2011). Thus, we fitted a reinforcement learning model-free algorithm, a reinforcement learning model-based algorithm and a hybrid account which assumes that choices at the first-level are generated based on the weighted combination of values from these two reinforcement learning models.

The task has three distinct states corresponding to the three sets of fractal stimuli: the first-stage state, $s_A$, and two second-stage states, $s_B$ and $s_C$. On each trial, $t$, subjects see a first-stage state, $s_{1,t}$ ($=s_A$), in which action $a_{1,t}$ is taken. This is followed by a second-stage state, $s_{2,t}$ (either $s_B$ or $s_C$) in which action $a_{2,t}$ is taken.

A model-free agent estimates a value function for each state-action pair. Thus, a prediction error, $\delta_{i,t}$, is computed and used to update value of the corresponding state-action: $Q_{MF}(s_{i,t}, a_{i}) \leftarrow Q_{MF}(s_{i,t}, a_{i}) + \alpha_{i} \delta_{i,t}$, where $\delta_{i,t} = r_{i,t} + Q_{MF}(s_{i+1,t}, a_{i+1,t}) - Q_{MF}(s_{i,t}, a_{i,t})$ is the prediction error at each stage and $\alpha_{i}$ is the learning rate parameter at either stage. Note that for first-stage choices, there is no direct reinforcement ($r_{1,t} = 0$) and for the second-stage choice, $Q_{MF}(s_{3,t}, a_{3,t}) = 0$, since there is no following state. The first-stage state-action value is also updated using an eligibility trace parameter, $\lambda$, to capture immediate effects of second-stage reinforcement on the first-stage state: $Q_{MF}(s_{1,t}, a_{1,t}) \leftarrow Q_{MF}(s_{1,t}, a_{1,t}) + \alpha_{1} \lambda \delta_{2,t}$.
A model-based agent takes into account transition probabilities to estimate the value of actions. Thus, this algorithm calculates the first-stage action based on the transition maps. Because the nature of the transition matrix (i.e. existence of rare and common transitions) is instructed, similar to Daw et al. (Daw et al., 2011), it is assumed that subjects choose between two possibilities: whether \( s_B \) is the second-stage set commonly associated with action \( a_A \) at first-stage, or vice versa that \( s_C \) is the one commonly associated with action \( a_A \) at first-stage. Without loss of generality, similar to Daw et al. (2011), we assume that probability of common and rare transitions are 0.7 and 0.3, respectively; if these are changed, other parameters of the model will rescale to give the same likelihood (Daw et al., 2011). Therefore, the model-based values of first-stage actions are computed as follows:

\[
Q_{MB}(s_A, a_j) = P(s_B|a_j)Q_{MF}(s_B, a_{max}) + P(s_C|a_j)Q_{MF}(s_C, a_{max}),
\]

where \( a_{max} \) is the action in the corresponding state that maximized \( Q_{MF} \) at the second-stage. Since the second-stage states are terminal states, model-based value of actions at the second-stage is assumed to be equal to that of model-free.

Finally, the hybrid account computed a weighted average of action value of model-based with that of model-free: \( Q_{hybrid} = wQ_{MB} + (1-w)Q_{MF} \), where \( 0 < w < 1 \) is a weight parameter. Higher values of \( w \) are associated with higher degree of model-based (and lower degree of model-free) influences on choice. For \( w = 0 \) and \( w = 1 \), the hybrid account is equivalent to pure model-free and pure model-based, respectively. A softmax transformation was then used to generate probability of choice for all models based on distinct decision noise parameters for each stage, \( \beta_i \), and a perseveration parameter, \( \phi \), which captures first-stage perseverance or switching tendency in choices regardless of action values (Lau and Glimcher, 2005).

**Model fitting and model selection**

We estimated parameters of each model separately for each participant using non-linear derivative-based optimization algorithm as implemented in fminunc tool in MATLAB (©Mathwork). All three models have second-stage learning rate, \( \alpha_2 \), two decision noise parameters, \( \beta_1 \) and \( \beta_2 \), and the perseveration parameter, \( \phi \). The model-free and hybrid accounts have two additional parameters for updating actions values at the first-stage, \( \alpha_1 \) and \( \lambda \). The hybrid model has one key weighting parameter, \( w \), for combining action-values of model-based and model-free at the first-stage. Four parameters, \( \alpha_1, \alpha_2, \lambda \) and \( w \) are bounded between 0 and 1. The decision noise parameters are bounded to positive values.
For bounded parameters of each model, we fitted parameters in the infinite real-space of Gaussian distribution parameter values and transformed them before feeding them into the models using appropriate transformation functions (sigmoid for parameters bounded between 0 and 1; exponential for parameters > 0). This method enabled us to employ unconstrained optimization techniques that are usually more robust than constrained ones. Similar methods have been adopted for fitting reinforcement learning models to choice data in this task (e.g. in Wunderlich et al., 2012b). A wide Gaussian prior, Normal(0,10), was assumed for all parameters (with zero mean and a broad variance of 10). Free parameters of each model were estimated to maximize log-likelihood of data plus log-prior (maximum a posteriori), where the likelihood is defined across both first-stage and second-stage choices, similar to previous works on this task (Daw et al., 2011). The prior distributions of parameters used in this study were broader than those of Daw et al. (2011).

We computed model-evidence for every model and every subject using Laplace approximation (MacKay, 2003), which penalizes complexity of the model by integrating out the free parameters. We then used the approximated model evidence to perform a random-effect Bayesian model comparison across all participants, a procedure which takes the model identity as random-, in contrast to fixed-, effect (Rigoux et al., 2014). We also used the approximated model evidence to compare models for each subject separately. For this analysis, a log-Bayes-factor>3 was considered as significant because the corresponding Bayes factor is more than 20 (cf. the classical p<.05 criterion). The log-Bayes-factor>2.3 was also considered as trend towards significance because it corresponds to p<0.1.

**Data acquisition and image processing**

Structural and diffusion images were collected using a 3-Tesla Siemens magnetic resonance imaging scanner. T1-weighted high resolution MP-RAGE structural image was collected (voxel size =1mm isotropic, GRAPPA acceleration factor 2). DTI scanning was performed with the following parameters: 64 slices interleaved acquisition mode (TE/TR = 89/6700 ms, flip angle = 90, FOV = 220 mm, voxel size = 2.2mm isotropic). DTI scans consisted of 7 scans without diffusion weighting (b = 0) and 61 scans with diffusion weighting (b = 1000 s/mm2) applied along the non-collinear directions.

All DTI preprocessing were conducted using FSL tools. Preprocessing of DTI data was performed based on the standard FSL protocol. BET was used to automatically extract brains from T1 (Smith, 2002) and images were manually checked for all samples and re-extracted if not successful. FNIRT was used for non-linear
registration of structural images to standard template (Jenkinson et al., 2012). Registered images were manually checked and FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002) was used for registration of structural images in four subjects where FNIRT was not successful. FDT was used to correct the DTI data for head movement and eddy current correction, brain extraction and tensor model fitting. The diffusion parameters were then sampled for each voxel using BEDPOSTX (Behrens et al., 2007a).

**Imaging analysis pipeline**

The goal of this study was to investigate the relationship between behavioural indices of model-based/model-free control quantified by the multistep decision task and anatomical circuitry connecting the striatum to the frontal cortex using DTI and probabilistic tractography. To quantify frontostriatal structural connectivity, we employed a fully automated procedure to compute connectivity maps between the striatum and the frontal cortex. To achieve this, we first performed a parcellation of frontal cortex based on its connectivity with the striatum. This analysis resulted in 5 clusters (see below for details). Next, connectivity between each striatal voxel and each of the five frontal clusters was computed. This resulted in 5 connectivity images per subject, quantifying connectivity between each striatal voxel and the 5 frontal clusters.

**Striatum-based parcellation of frontal cortex**

First, we created a striatal mask in MNI space using the Harvard-Oxford subcortical atlas. The MNI frontal lobe mask was used for frontal cortex. For computational feasibility, the frontal mask was re-sampled to 4mm isotropic voxel size. These masks were then transformed to each participant's native diffusion space using registration wrap images and matrices computed during preprocessing. Probabilistic tractography was then performed in native diffusion space using PROBTRACX (Behrens et al., 2007a), where tracts seeded from every voxel within the frontal lobe and its connectivity with all striatal voxels was quantified in each participant (Behrens et al., 2007a). This procedure computes a connectivity matrix, which characterizes every voxel within the frontal lobe based on its connectivity pattern with striatal voxels. The connectivity matrix was used to generate a symmetric cross-correlation matrix, which reflects the correlation in connectivity fingerprint of frontal voxels. This cross-correlation matrix was then subjected to K-means clustering, a well-known algorithm for clustering used previously for parcellation of brain regions (Beckmann et al., 2009; Mars et al., 2011; Neubert et al., 2014; Piray et al., 2015), to identify voxels sharing similar striatum-connectivity profiles. Since the correct number of frontal clusters is unknown, we performed a stability analysis to identify the most consistent and coherent number of clusters (see below.
Subjects were randomly divided into two groups and a series of parcellation into 2 to 8 clusters was carried out separately for each group. The clustering solutions based on data from two groups were then compared to examine their consistency as a function of number of clusters. This procedure was repeated for 100 randomly division of subjects to two groups and used to obtain a stability index. Tractography was performed separately for the right and left hemispheres.

**Stability analysis of parcellation solution**

To ensure that the parcellation scheme is robust at the between-subject level, we performed a stability analysis, which identifies the largest number of clusters resulting in a significantly robust clustering solution. To achieve this, we assessed whether two clustering solutions calculated based on two independent datasets (e.g. by dividing subjects randomly to two groups) were matched. Here, we provide a mathematical explanation of our approach (copied from Appendix 1 in Piray et al. (2015)):

Two sets of clusters (A and B, each with K clusters) were defined as matched based on the following criteria: First, for every cluster in A and every cluster in B, an overlap index was defined, which corresponds to the number of voxels that overlap between the two clusters. Specifically, for every cluster \( a_i \) in A and every cluster \( b_j \) in B, the overlap index was defined as \( N_{i,j} / \min(N_i, N_j) \), where \( N_i \), \( N_j \) and \( N_{i,j} \) are the number of voxels in \( a_i \), \( b_j \) and their intersection, respectively. Next, for every cluster \( a_i \) in A, \( b_j \) in B was defined as matched if it had the largest overlap index with \( a_i \). Finally, A and B were considered as matched if each cluster in A was matched with one and only one cluster in B; and vice versa if each cluster in B was matched with one and only one cluster in A. This procedure also gives a one-to-one mapping between “labels” of clusters in A and B, regardless of anatomical location of voxels.

**Connectivity maps between the striatum and frontal clusters**

Having established the target frontal regions, probabilistic tractography (using PROBTRACKX tool in FSL) was seeded from each voxel in the striatum with the five identified clusters as targets (using the classification mask option in PROBTRACKX). This procedure created five images, one for each frontal target, of probability values where each voxel value corresponds to the number of pathways that begins at that voxel and ends in the target region. Tractography was performed separately for the right and left hemispheres. All maps were smoothed with a 6mm Gaussian kernel.
We also performed an analysis to make inference on the anatomical directionality of tracts, which masks out those tracts passing through the thalamus. For this analysis, the JHU atlas was used to create a mask of the anterior limb of the internal capsule (Oishi et al., 2010). We used this mask as an exclusion mask and re-performed probabilistic tractography analysis to assess connectivity between the striatum and the frontal clusters. Therefore, this analysis simulates a lesion in anterior limb of the internal capsule, thereby discarding all fibers running from the striatum to frontal lobe along the striatal-thalamo-cortical pathway.

**Statistical analysis**

We then investigated whether behavioral indices of model-based and model-free control could be predicted by frontostriatal connectivity maps computed in the previous steps. Since tract strength values are non-normally distributed, non-parametric analysis (rank correlation) was performed using tools from FSL software (FSL Randomize with 5000 permutation tests) (Winkler et al., 2014). Threshold-free cluster enhancement (TFCE), as implemented in FSL (Smith and Nichols, 2009), was used to boost signal in areas that exhibit spatial clustering (with variance smoothing kernel of 6). All resulting statistical maps were corrected (p<.05) at the voxel level, separately for the left and right striatum, for family-wise error due to multi-voxel comparisons. All reported coordinates are the MNI coordinates.

**Results**

**Behavioral data**

The critical feature of the multistep decision task is the probabilistic nature of the transition from the first to the second-stage set. Each first-stage choice led predominantly (70%) to one of the two second-stage sets (common transition), and less frequently (30%) to the other set (rare transition) (Figure 1A). Model-based and model-free accounts make different predictions about participants’ choices in rare-transition trials. A model-based system reinforces the first-stage choice predominantly associated with the rewarded second-stage choice, which results in decreasing the probability of choosing the first-stage action that is ultimately rewarded after rare transitions (Figure 1B, left). In contrast, a model-free system is blind to transition probabilities and therefore reinforces those first-stage choices ultimately rewarded regardless of the transition (Figure 1B, middle). Therefore, one can model the probability of repeating the first stage choice on the subsequent trial (stay probability) as a function of two key events on the current trial. The two key events are whether or not reward was delivered, and whether or not the transition was common or rare. Model-free and model-based components of
Figure 1 Task setup, model predictions, and behavioral data. A) Task setup: Participants chose between two fractal stimuli, which led probabilistically to one of the two different second-stage sets. B) Model predictions and observed behavior. Left: if choice were completely controlled by the model-based system, the first-stage choice predominantly associated with the rewarded second-stage choice would be reinforced; Middle: if choice were completely controlled by the model-free system, then repeating the first-stage choice in the subsequent trial (stay-probability) is a function of reward delivery regardless of the transition occurred. Right: data averaged over all subjects shows signature of both systems. The analysis of stay probability data revealed a significant main effect of reward delivery (i.e. model-free signature) as well as an interaction between reward delivery and transition (i.e. model-based signature). C) Individual variability in the reliance on the model-based system. Subjects are sorted in descending order based on reward-by-transition interaction effect, which is an index of model-based control in the task. In half of participants, the hallmark of model-based control is clearly observable. However the other half of participants shows no evidence of reliance on the model-based strategy. Insets: Mean stay probabilities as a function of reward and transition. Bottom left inset plot shows the data from the median-split half of individuals with a large reward-by-transition effect; the top right inset plot shows the data from the median split half of individuals with a small reward-by-transition effect. Error-bars reflect standard error of the mean.
behavior could then be quantified as the main effect of reward and the interaction effect of reward and transition, respectively.

Across participants, the presence of reward increased the probability of repeating the first-stage choice (main effect of reward, $F(1,30)=28.53$, $p<0.001$), an indication that model-free control influenced participants’ choices (Figure 1B). Additionally, the type of transition also affected first-stage choices (reward-by-transition interaction, $F(1,30)=8.29$, $p=0.007$), an indication that model-based control also influenced participants’ choices (Figure 1B). There was no main effect of transition on choice ($F(1,30)=0.21$, $p=0.65$), as predicted by both model-based and model-free accounts. There was a significant positive intercept ($F(1,30)=77.14$, $p<0.001$), indicating a tendency to implement the choice made on the previous trial regardless of reward delivery and transition (Lau and Glimcher, 2005). Table 1 summarizes the result of this analysis.

Next, we elaborated on this factorial group-level analysis by considering the whole history of rewards obtained prior to a given trial, and by considering individual-level data. This was achieved with a Bayesian model selection procedure comparing the fit of the behavioral data to the predictions of three different models. The first model was a hybrid reinforcement learning model previously used to account for
choices in this task (Daw et al., 2011). The hybrid model combines learned values of model-based and model-free strategies on a trial-by-trial basis and employs their combination for action selection. The other two models were pure model-free and pure model-based accounts. Across the group, random-effect Bayesian model selection (Rigoux et al., 2014) indicated that the hybrid account provides the most parsimonious model given the population-level data (exceedance probability of 1.0, expected posterior model probability of 0.94; Table 2). At the individual level, pair-wise comparison between the hybrid and model-based accounts revealed that the hybrid account significantly outperformed the model-based account in 31 out of 31 participants (log-Bayes-factor>3.0, Table 2), while a similar pair-wise comparison between the hybrid and the model-free account revealed that hybrid outperformed model-free account only in 6 out of 31 participants (log-Bayes-factor>3.0). The latter finding is not driven by a particular statistical threshold: relaxing the log-Bayes-factor to 2.3 (corresponding to p<0.1 in frequentist statistics) leads to the hybrid account providing a better fit than the model-free account in 12 out of 31 participants. The finding is also graphically confirmed by ranking participants according to their reward-by-transition interaction effect in the factorial analysis: whereas the signature of the model-based strategy was not evident in half of subjects, it was clearly seen in the other half (Figure 1C). These findings suggest that the participants consistently used model-free control, whereas the use of model-based control varied across the sample.

Table 1 Logistic regression analysis of behavioral data. Mean estimate of regression coefficients and their standard error (SE) are shown (arbitrary unit).

<table>
<thead>
<tr>
<th>Effects</th>
<th>Estimate (SE)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reward</td>
<td>0.32 (0.06)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Transition</td>
<td>-0.03 (0.06)</td>
<td>0.65</td>
</tr>
<tr>
<td>Reward x transition</td>
<td>0.24 (0.08)</td>
<td>0.007</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.31 (0.15)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

P-values of effects across group are reported. This analysis indicates a significant effect of the reward of previous trial and an interaction between reward and transition of previous trial on stay probability on the current trial.
Further quantitative analyses confirmed the presence of large individual differences in the use of model-based control in this task. Namely, the reward-by-transition interaction values are not normally distributed across the sample \((p=0.017,\text{ Lilliefors test})\), despite its relatively large size \((n=31)\).

Similar set of analyses revealed individuals exhibit less variability in employing model-free control. First, model fits showed that all subjects employed model-free strategy, as the hybrid account outperformed the pure model-free account in all 31 participants significantly (Table 2). Furthermore, splitting the sample by the median value of reward effect shows that model-free deployment was significantly observable even in that half of subjects who employed model-free strategy less than the other half \((F(1,15)=6.75, p=0.02)\). Finally, and in contrast to the reward-by-transition interaction effect, no evidence in favor of non-normal distribution of reward effect was found across the sample (Lilliefors test, \(p>0.05\)).

**Model-based correlation with striatal anatomical connectivity**

DTI data was used to define a connectivity matrix between the striatum and frontal cortex in each participant, in order to test whether their structural connectivity predicts individual differences in employing model-based control. The connectivity matrix was then used to parcellate frontal cortex by identifying voxels with a shared profile of connectivity with the striatum (Beckmann et al., 2009; Mars et al., 2011; Neubert et al., 2014). A stability analysis was performed to identify the most consistent and coherent number of clusters. This stability analysis revealed that five clusters could be identified reliably at the group level in both hemispheres (Monte Carlo randomization test, \(p<0.001\)). In addition, although this parcellation scheme was blind to voxel location, voxels clustered into five

<table>
<thead>
<tr>
<th>Model</th>
<th>No. free parameters</th>
<th>Exceedance probability</th>
<th>Expected posterior</th>
<th>No. favoring hybrid with (\text{LBF}&gt;3.0)</th>
<th>No. favoring hybrid with (\text{LBF}&gt;2.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid</td>
<td>7</td>
<td>1.0</td>
<td>0.94</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pure model-free</td>
<td>6</td>
<td>0.0</td>
<td>0.03</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Pure model-based</td>
<td>4</td>
<td>0.0</td>
<td>0.03</td>
<td>31</td>
<td>31</td>
</tr>
</tbody>
</table>

The hybrid model outperforms both pure model-based and pure model-free accounts based on random-effects Bayesian model comparison results as shown by both exceedance probability and expected posterior probability across models. However, large individual differences in deployment of model-based control are evident, as the hybrid account outperformed the pure model-free account only in 6 subjects with log-Bayes-factor of 3.0 (cf. \(p<0.05\)). Even for log-Bayes-factor of 2.3 (cf. \(p<0.1\)), the hybrid model outperformed the pure model-free account only in 12 subjects.

Further quantitative analyses confirmed the presence of large individual differences in the use of model-based control in this task. Namely, the reward-by-transition interaction values are not normally distributed across the sample \((p=0.017,\text{ Lilliefors test})\), despite its relatively large size \((n=31)\).

Table 2 Bayesian model comparison.
anatomically coherent parcels, which were largely symmetric across both hemispheres (Figure 2). There are additional medio-lateral subdivisions within each of the five clusters when cytoarchitecture and cortico-cortical connections are considered (Beckmann et al., 2009; Sallet et al., 2013; Neubert et al., 2014). However, since the parcellation scheme only considered fronto-striatal connectivity, the frontal clusters should be interpreted as cortical territories that are homogeneous from a striatal point of view, given DTI data.

Figure 2 Connectivity-based parcellation of frontal cortex resulted in 5 distinct clusters consistently identified across participants. These clusters should be interpreted as frontal regions segregated according to their striatal connectivity profile. The data-driven parcellation (clustering procedure) was blind to the anatomical location of the frontal voxel, yet those voxels clustered into five anatomically coherent territories. This procedure resulted in a map with anteroventral to posterodorsal gradient organized in accordance with known profiles of frontostriatal connectivity (Draganski et al., 2008b). The map consisted of five clusters: a precentral cluster (in black); a posterior prefrontal cluster (in brown); a dorsal prefrontal cluster (in red); an anterior prefrontal cluster (in orange); and a ventromedial prefrontal cluster (in yellow).
The parcellation procedure resulted in a map with anteroventral to posterodorsal gradient organized in accordance with known bands of frontostriatal connectivity (Draganski et al., 2008b; Cohen et al., 2009; Haber and Knutson, 2010). The five clusters included: i) a precentral cluster overlapping with motor areas of the frontal lobe such as frontal operculum cortex and precentral gyrus; ii) a posterior prefrontal cluster including pre-supplementary motor area and posterior parts of superior- and middle- frontal gyrus; iii) a dorsal prefrontal cluster including a large portion of inferior frontal gyrus and anterior parts of middle- and superior-frontal gyrus. This dorsal prefrontal cluster also overlapped with posterior parts of anterior cingulate gyrus and paracingulate gyrus; iv) an anterior prefrontal cluster including the most anterior part of the paracingulate and anterior cingulate gyrus as well as dorsal parts of frontal pole; v) a vmPFC cluster including frontal orbital cortex and ventral parts of frontal pole.

The degree of model-based strategy deployment, quantified in each participant as the reward-by-transition interaction effect, was significantly associated with the strength of connectivity between the vmPFC cluster and the medial striatum (p<0.05, FWE corrected; Figure 3A; local maximum within the left striatum x=-20, y=6, z=-6, local maximum within the right striatum x=20, y=2, z=-6). Individuals relying more on model-based control had stronger structural connectivity between the vmPFC cluster and the medial striatum. This effect was anatomically specific: No significant correlation was found between model-based control and striatal connectivity with the other frontal clusters. Furthermore, the effect was not driven by strong between-clusters inhomogeneities in connectivity variance: The maximum standard deviations across all striatal voxels for each map were comparable, with the anterior prefrontal cluster, the dorsal prefrontal cluster and the posterior prefrontal cluster showing larger variability across participants than the vmPFC cluster.

Similar results were obtained when the degree of model-based strategy deployment was indexed with the weighting parameter, $w$, of the hybrid model (Figure 3C). Higher values of $w$, corresponding to higher degree of model-based control, are associated with stronger connectivity between vmPFC and medial striatum (significant in the left striatum, p<0.05, FWE-corrected; local maximum x=-26, y=-8, z=-4). This was expected, as the weighting parameter was strongly correlated with the degree of model-based quantified as reward-by-transition interaction effect in the factorial model (r=0.64, p=0.0001).
DTI does not provide directional information, but the anatomical organization of the fronto-striatal circuits allows one to examine whether the effect described above is driven by direct projections from vmPFC to medial striatum, or by thalamus-mediated connections from medial striatum to vmPFC. Accordingly, we performed another tractography analysis, by masking out tracts passing through the thalamus, to make inference on the anatomical directionality of the effects. This analysis revealed effects similar to those reported above (Figure 3B, p<0.05, FWE corrected).

**Figure 3** Individual differences in model-based control were predicted by anatomical connectivity strength between the vmPFC cluster and the striatum. Maps represent pseudo t-statistics from non-parametric correlation analysis. All maps are thresholded at p<0.05, familywise error corrected. **A** Connectivity between the vmPFC cluster and the bilateral striatum predicts individual differences in model-based control, as indexed by reward-by-transition effect shown in Figure 1C. The effects are present in the medial striatum, particularly the left medial caudate nucleus and bilateral dorsal anterior putamen. **B** The same effects are present even when those tracts running from the striatum to the frontal cortex through thalamus were discarded, suggesting that effects are, at least partly, driven by top-down afferences from the vmPFC to the striatum. **C** Similar correlation was found when we considered the weight parameter, w, of the hybrid model as the degree of model-based influences. **D** Rank scatter-plot from panel A, separately for the left and right hemisphere tracts, averaged over all voxels showing significant correlation.
corrected; local maximum within the left striatum $x=-20, y=6, z=-6$; local maximum within the right striatum $x=24, y=2, z=-8$), suggesting that those effects are largely driven by top-down afferences from the vmPFC to the medial striatum. The complementary control analysis, seeding tractography from the anterior limb of internal capsule while excluding all striatal voxels, did not reveal significant effect even at a very lenient statistical threshold ($p<0.1$ uncorrected for multiple comparisons). This control analysis provides a complementary, although negative proof that the effects of cortico-striatal connectivity on model-based control are driven by top-down connections from the vmPFC to the striatum.

We also performed a similar analysis to assess whether individual differences in model-free deployment, quantified as the main effect of reward in the task, could be predicted by the strength of connectivity between the striatum and the precentral/posterior prefrontal clusters. There was no significant correlation between the magnitude of model-free control and the strength of the connectivity between those frontal clusters with the striatum. A post-hoc analysis extending this approach to the remaining frontal clusters revealed a significant negative correlation between right medial caudate nucleus and the dorsal prefrontal cluster ($p<.05$, FWE-corrected; local maximum $x=10, y=8, z=2$). Individuals with a higher degree of model-free strategy deployment had lower structural connectivity between the right dorsal prefrontal and the right medial caudate nucleus.

Based on animal and human literature on goal-directed and habitual behavioral control, we hypothesized that connectivity between the frontal cortex and the striatum predicts individual differences in model-based control. However, recent studies have suggested that there are other regions implicated in model-based control. Specifically, it has been hypothesized that model-based control might implicate the amygdala, hippocampus, lateral prefrontal cortex and/or the default model network (Doll et al., 2012; Daw and Dayan, 2014; Dayan and Berridge, 2014). Therefore, we performed an exploratory analysis to test whether the connectivity between the vmPFC cluster and these regions are correlated with the degree of model-based control. These regions were defined according to Harvard-Oxford atlas, except the lateral prefrontal cortex which is defined according to diffusion-based connectivity-parcellation of human dorsal prefrontal cortex (cluster 6 in (Sallet et al., 2013). These atlases are available in FSL.

We found marginal effects in a few voxels in the left posterior cingulate cortex, a hub of the default mode network. The connectivity between vmPFC and the left posterior cingulate was positively associated with the degree of model-based control ($FWE<0.05$; peak at $x=-4, y=-41, z=38$, corrected p-value in peak, $p=0.048$).
Model-based association with white matter bundles

Probabilistic tractography estimates the probability distribution of the parameters of a crossing fiber model of diffusion magnetic resonance imaging data. The tensor model is a simpler model of diffusion magnetic resonance imaging (Basser et al., 1994), which provides a scalar measure, referred to as fractional anisotropy, that has been related to white matter microstructure integrity (Song et al., 2003). Here, we employ tract-based spatial statistics (Smith et al., 2006) to test whether the association between model-based behavior and vmPFC tract strength, as revealed by probabilistic tractography, is accompanied by an association between model-based behavior and tract integrity, as quantified using fractional anisotropy. To this end, we performed voxel-wise correlation analyses of the skeletonised fractional anisotropy data, focusing on four major white matter bundles shown to carry tracts originating from the vmPFC (Lehman et al., 2011; Jbabdi et al., 2013): the uncinate fascicle, the corpus callosum, the superior longitudinal fascicle and the cingulum bundle. All these masks were created based on JHU white-matter atlases (Wakana et al., 2007; Hua et al., 2008). This analysis revealed a significant correlation between tract integrity in the cingulum bundle and the degree of model-based control (p<0.05, family-wise error corrected; local maximum x= -19, y= -36, z= +34). No significant correlation was found in other masks.

However, the interpretability of results obtained using the tensor model of diffusion data in regions with crossing-fibers has been questioned by many authors (Tournier et al., 2004; Parker and Alexander, 2005; Behrens et al., 2007a; Jbabdi et al., 2010). One solution to this issue is to use tract-based spatial statistics with measurements from models dissociating different fibers in different directions (Jbabdi et al., 2010), such as bedpostX (Behrens et al., 2007a). Therefore, we repeated the above analysis with partial volume fraction values estimated along with the first fiber orientation quantified by bedpostX. We found very similar results, with highly significant correlation between tract integrity voxels in the cingulum bundle and model-based scores (p<0.05, family-wise error corrected; Figure 4A; local maximum x=-7, y= +5, z = +32), but not in other masks. These effects survived correction for comparison in multiple masks too. Thus, participants with higher tract integrity in the cingulum bundle showed higher degree of model-based behavior in the task (Figure 4B).

One question raised by this analysis is whether the brain-behavior correlation with tracts connecting the vmPFC with the striatum (Figure 3) is mediated by tracts passing through the cingulum bundle. To assess this, we repeated our original probabilistic tractography analysis of connectivity between the striatum and the vmPFC cluster and used the cingulum bundle as an inclusion mask. This
analysis discards all the tracts do not pass through the cingulum bundle. We found that the strength of tracts between vmPFC and a dorsomedial striatal region, passing through the cingulum bundle, is significantly associated with the degree of model-based control (p<0.05, FWE corrected; Figure 4C; local maximum within the left striatum x=-13, y=14, z=-6, local maximum within the right striatum x=8, y=8, z=-4).

Following a reviewer’s comment, we have also performed voxel-based morphometry (Ashburner and Friston, 2000) analysis to assess whether individual variability in
model-based control is also associated with individual variability in grey matter density in the vmPFC cluster and/or the striatum, using tools implemented in SPM8 software (Ashburner and Friston, 2000; Ashburner, 2007). Whole-brain analysis revealed no significant association, even at the lenient threshold of p<0.001 uncorrected. Further region-of-interest analyses in the vmPFC and the striatum revealed no significant correlation either at the voxel-level (family-wise error corrected, p<0.05) or at the cluster-level (not even when we used p<0.01 as uncorrected p-value for cluster-level inference). These analyses suggest that the correlation between model-based control and vmPFC-striatum tract strength is not accompanied by a similar correlation with grey matter density.

Discussion

In this study, we tested the hypothesis that the relative contribution of model-based and model-free control systems to decision making depends on the relative strength of anatomical connectivity within frontostriatal circuits involved in learning and behavioral control. We exploited the presence of large and systematic inter-individual differences in the use of model-based control during instrumental actions (Figure 1, (Daw et al., 2011)). This study shows that the use of model-based control is predicted by neuroanatomical differences in the structural coherence of white-matter tracts from the vmPFC to the medial striatum. The finding indicates that individuals with more coherent afferences from vmPFC to medial striatum are more likely to rely on a model-based system to control their instrumental actions. Furthermore, an analysis based on making out of bottom-up tracts suggests that this effect is driven by top-down influences from vmPFC to medial striatum. These findings extend and qualify previous knowledge on how the control of goal-directed behaviour is neuronally implemented through the vmPFC-striatal circuitry (Balleine and Dickinson, 1998; Corbit and Balleine, 2003; Yin et al., 2005; Valentin et al., 2007; Tanaka et al., 2008; Gläscher et al., 2009).

Ventromedial prefrontal and striatal contributions to goal-directed behavior

Previous work has suggested that, when goal-directed and habit-based control compete, model-based and model-free strategies are computed in the caudate nucleus and in the posterior putamen, respectively, while the vmPFC integrates those computations (Wunderlich et al., 2012a). The pattern of behavioral and cerebral inter-individual differences observed in the present study suggests a different neurocognitive architecture. The present findings show that the vmPFC biases the balance between model-free and model-based control. The bias is
implemented by modulating participants’ reliance on model-based control through cortico-striatal projections from the vmPFC to the caudate nucleus. This architecture fits well with a recent hierarchical model of action control, in which shifting from model-free to model-based control is itself a goal-directed decision controlled by a model-based system (Dezfouli and Balleine, 2013; Daw and Dayan, 2014). For instance, the pattern of vmPFC activity reported in Wunderlich et al. (2012a) could reflect the implementation of goal-directed choices between performing overtrained stimulus-response associations (presumably model-free) and navigating a complex decision-tree (presumably model-based). The present findings also fit with the notion that vmPFC contributes to decision making by encoding an abstract, cognitive map of task space (Wilson et al., 2014). The multi-step decision task used here is designed to make participants choose between options followed by unobservable probabilistic transitions between states (Daw et al., 2011). By providing an explicit computational account on how those choices are biased towards model-free or model-based control systems, the present study extends previous reports linking vmPFC-caudate nucleus connectivity to flexible goal-directed control (de Wit et al., 2012). In that study, “slips of actions” were used to quantify habitual responses, but this behavioral outcome does not precisely capture the relative balance between model-based and model-free control (Dolan and Dayan, 2013). Here, we show that the vmPFC biases the relative contribution of model-based and model-free systems, as implemented in the caudate nucleus, on the basis of a cognitive map of task-space.

Model-based control has previously been shown to vary with state factors, such as stress (Otto et al., 2013b), working memory capacity (Otto et al., 2013a) and dopamine synthesis capacity in the striatum (Deserno et al., 2015). In prior work, we have shown that fronto-striatal tract strength can predict dopamine’s effect on cognitive control and frontostriatal functioning (van Schouwenburg et al., 2013). Accordingly, it is possible that the correlation between model-based control and individual differences in the strength of the vmPFC-striatum tract, observed here, reflects differential sensitivity to dopamine-related states, such as stress and working memory. Another indirect evidence comes from studies showing that vmPFC response to reward is related to state stress levels (Treadway et al., 2013), and studies showing that prefrontal-dorsomedial striatal structural connectivity, measured using DTI, predicts individual differences in reward dependence (Cohen et al., 2009). This hypothesis can be tested in future studies, combining DTI with an interventional (psychopharmacological or stress-induction) approach.
Implications for psychiatric disorders
Disruption of the balance between goal-directed and habitual modes of behavioral control might account for several impulsive/compulsive psychiatric disorders, such as impulse control disorders, obsessive-compulsive disorders, obesity, and drug addiction (Brewer and Potenza, 2008; Belin et al., 2013; Smith and Robbins, 2013; Gillan and Robbins, 2014). For instance, it has been recently shown that compulsive disorders are associated with a bias towards model-free control, at the expenses of reduced model-based control (Voon et al., 2014). The present findings raise the possibility that this pathological bias might be mechanistically implemented through altered anatomical connectivity between vmPFC and the caudate nucleus.

Interpretational issues
In this study, we exploited the presence of large individual differences in model-based control and investigated whether these differences could be predicted by neuroanatomical differences in fronto-striatal circuitry. This approach builds on previous reports showing that subjects’ behavior is stable across repetitions of this task (Wunderlich et al., 2012b; Smittenaar et al., 2013). For example, Wunderlich and colleagues conducted a within-subject study in which subjects received levodopa and placebo in two sessions and were tested in the same paradigm used in this study (Wunderlich et al., 2012b). They found no evidence in favor of different performance, either in stay probability or parameter fits, across sessions. Similar observations have been reported in other within-subject studies that used the same multistep decision paradigm (Smittenaar et al., 2013).

The multistep decision task used in this study manipulated the value of actions, while the transition probabilities of the task were fixed. Therefore, it was not possible to dissociate two important aspects of model-based control, namely learning the value of the task actions and learning a model of the task environment. In the present study, a post-hoc analysis revealed that structural connectivity between the right dorsolateral prefrontal cortex and the right caudate nucleus is negatively correlated with reliance on model-free control. Accordingly, it has been shown that interference with the same portion of the right dorsolateral prefrontal cortex shifts the balance of the two systems towards model-based control (Smittenaar et al., 2013). Future studies challenging participants to learn multiple models of the task environment might be able to expand on the notion that the dorsolateral prefrontal cortex is associated with learning probabilities of state transitions (Gläscher et al., 2010), and show how this region interacts with the vmPFC-caudate circuit when goal-directed model-based actions are generated.
It might be argued that this study failed to isolate a structural counterpart to participants’ reliance on habits, despite evidence linking structural connectivity between posterior putamen and premotor cortex to habitual responses (de Wit et al., 2012). In fact, there are important differences between the habitual responses considered by de Wit et al. (2012), and the model-free actions elicited by the current multistep decision task. In contrast to habitual “slips of actions”, the current model-free actions remain sensitive to reinforcements, but are blind to architecture of states in the environment. Furthermore, the responses performed in the multistep decision task had no consistent spatial mapping, as choices were randomized across trials. It remains to be seen whether other forms of model-free learning, such as action-sequence learning directly linking stimuli to sequences of actions (Dezfouli et al., 2014), might be suitable for capturing habitual responses.

There are important anatomical and functional differences between lateral and medial portions of each of the five frontal clusters considered in this study (Rushworth et al., 2011, 2012). Future studies might be able to test whether and how those differences, largely determined on the basis of cytoarchitectonic features and cortico-cortical connectivity, are also relevant for understanding the relation between fronto-striatal connectivity and model-based control.

**Conclusion**

This study investigated neural sources of individual differences in the computational bases of human choice by linking parameters of a normative learning model to structural cerebral features. The evidence indicates that a circuit connecting vmPFC to the medial striatum predicts inter-individual differences in participants’ reliance on model-based control. Individuals with stronger afferences from vmPFC to medial striatum are more likely to rely on a model-based system when controlling their instrumental actions. Explaining inter-individual variability in model-based decisions open the way to provide a mechanistic understanding of pathological computational processes associated with deficits in the balance between the goal-directed and habitual action control (Belin et al., 2013; Voon et al., 2014).
CHAPTER

5

Seeing an angry face impairs learning the value of information in an uncertain world

This chapter is adapted from:
Abstract

Optimal decision-making weights information on the outcome of a choice according to its uncertainty, which is a function of environmental volatility. Here we test whether emotional cues influence experienced volatility in the environment. Human participants learned contingencies between visual cues, responses, and financial outcomes, while the emotional content of the cues and the rate of change of the contingencies were independently manipulated. Participants learned the task contingencies and distinguished between stochastic and systematic changes in the environment, as quantified with a Bayesian hierarchical learning model. Participants’ learning rates and estimated volatility are lower when actions are cued by an angry face than by a happy face. At the neural level, emotions bias associative learning by modulating activity within anterior cingulate cortex and its connectivity with the dorsomedial striatum. Volatility learning signals, encoded in the dorsolateral prefrontal cortex, are disrupted after seeing angry face cues. These findings characterize the computations and neural circuits influenced by emotion during associative learning.
Introduction

Statistical regularities in the causal structure of the environment influence associative learning and choice (Yu and Dayan, 2005; Courville et al., 2006; Behrens et al., 2007b, 2008; Nassar et al., 2010; Payzan-LeNestour et al., 2013). To date, most formal learning models have focused on learning cue-action-outcome contingencies. For instance, the Pearce-Hall model (Pearce and Hall, 1980) proposes that an agent tracks associability, a time-varying metric of the degree to which a cue has previously resulted in a surprising outcome. The Pearce-Hall model can account for a number of biases, as those induced by emotions, by changing cue associability. However, accurate predictions and outcome maximization often require learning higher order statistics of the environment, such as volatility, a time-varying metric of systematic changes in outcome probability (Behrens et al., 2007b). Hierarchical Bayesian theories provide formal learning models that are able to dissociate between stochastic and systematic changes in the environment, but it remains unclear whether, and at which level, emotional biases influence those hierarchical structures. This human functional magnetic resonance imaging (fMRI) study uses a hierarchical Bayesian learning model to understand which computations and which neuronal mechanisms are influenced by emotion during associative learning.

According to hierarchical Bayesian learning theories (Behrens et al., 2007b; Iglesias et al., 2013), our decisions are guided not only by the outcomes of similar decisions made in the past, but also by the degree to which we perceive the world as stable or volatile. In these models (Behrens et al., 2007b; Mathys et al., 2011), the causal structure of the world is inferred by constructing and updating a hierarchical model of its sensory inputs, and action values are updated by the product of two constructs: i) prediction error, representing the difference between observed outcome feedbacks, such as reward or punishment, and current estimate of the likelihood to receive such feedbacks; ii) learning rate, representing uncertainty about current estimation of choice value. In a stable environment, outcome feedback is relevant and valuable for decision making when the agent is not certain about estimated action value, e.g. on first exposure to that environment. As observations accumulate and action value estimation improves, outcome feedback becomes less valuable. However, if the environment changes, e.g. an action-outcome contingency is reversed, outcome feedback becomes valuable again, as previous action value estimates do not hold anymore. Hierarchical Bayesian learning keeps track of environment volatility, and estimates it according to estimated uncertainty and noise in the environment. It has been shown that, in these accounts, updating action values depend on estimated volatility, and that activity in the anterior
cingulate cortex correlates with estimated volatility (Behrens et al., 2007b, 2008). However, it remains unclear how emotions modulate associative learning, namely whether emotions influence action-value learning directly or by biasing volatility estimation.

This study assesses the behavioral and cerebral effects of emotions on volatility tracking. We test whether emotions influence action value learning by modulating brain regions known to compute volatility (e.g. the anterior cingulate cortex (Behrens et al., 2007b, 2008)), or regions known to compute action value learning (e.g. the striatum (Delgado et al., 2000; Samejima et al., 2005; Schönberg et al., 2007)). We experimentally manipulate environmental volatility and we employ a hierarchical Bayesian modeling approach to quantify participants’ volatility estimates (Behrens et al., 2007b; Mathys et al., 2011). As reinforcement learning theories suggest that neural substrates of learning might depend on outcome valence (Daw et al., 2002; Cools et al., 2011), we consider both appetitive and aversive emotional cues (happy/angry faces) and outcomes (monetary gains/losses), in a 2x2 factorial design that distinguish the effects of emotion valence and outcome valence.

Materials and Methods

Participants
Forty-five female volunteers gave written informed consent approved by the local ethical committee (“Comissie Mensgebonden Onderzoek” Arnhem-Nijmegen) and participated in the study. Exclusion criteria were claustrophobia, neurological, cardiovascular or psychiatric disorders, regular use of medication or psychotropic drugs, heavy smoking and metal parts in the body. One participant did not finish the experiment due to headache. Data from all other forty-four participants were analyzed (all right-handed, mean age of 20.7).

Experimental design
We used data from a previously published (Ly et al., 2014) probabilistic learning task focused on the association between emotional choice biases and individual differences in social avoidance. Each participant completed 480 trials of a probabilistic learning task in the scanner. Each trial started with a face cue (happy or angry) presented on a color frame indicating the type of outcome valence (reward or punishment) at the end of the trial. Thus, there were four trial-types in a 2x2 factorial design with factors emotion (happy or angry) and valence (reward or punishment). There were 120 trials per trial-type. Participants were instructed
that the combination of emotional content of the face cue and color frame distinguished the four trial-types and that they had to learn the optimal response for each of the four cue-types separately. The response-outcome contingency was probabilistic and independent for each trial-type. The response-outcome contingency was reversed several times for each trial type, resulting in different degree of volatility in the course of experiment, while remaining counterbalanced across trial types. Specifically, each participant completed three sessions, with a 1-min break in between the sessions. Each session consisted of 160 trials, with 40 trials per trial-type. For each trial-type within a session, the probability of a positive outcome given a go-response could take one of the following combinations in two consecutive blocks: (i) 0.5, 0.2, 0.5, 0.2; (ii) 0.5, 0.2, 0.5, 0.8; (iii) 0.5, 0.8, 0.5, 0.8, where each session was associated with one of these combinations. The blocks with probability of 0.5 were short blocks with average length of 5 trials, and other blocks were long blocks with average length of 15 trials.

Emotional stimuli were adult Caucasian faces from 36 models (18 men) taken from several databases (Ekman and Friesen, 1976; Matsumoto and Ekman, 1988; Lundqvist et al., 1998; Martinez, and Benavente, 1998). Model faces were trimmed to exclude influence from hair and non-facial contours (van Peer et al., 2007; Roelofs et al., 2009). Model identity was counterbalanced, such that the model occurred equally often for each trial-type. The color frame (yellow or grey) indicating the possibility of reward or punishment was also counterbalanced across participants. On each trial, one of the face cues was presented centrally. Participants were then allowed to make a response 100 ms after cue onset, where they were required to make either a go- or a no-go-response within 1000 ms. If no response was made within 1000 ms, then a no-go-response was recorded. After a response-outcome delay of maximally 2000 ms (depending on the response time), the outcome was presented for 1000 ms (+10 cents for reward, -10 cents for punishment, and 0 cents for omitted reward or avoided punishment). The inter-trial interval was jittered (2500 to 4500 ms).

The relatively long time window for responding (1000 ms) ensured that no-go responses are not due to failure in making a go response. To illustrate this point, we tested each participant response-time separately for go-responses in every trial-type. This test revealed that for all participants and all trial types, response-time are significantly lower than 1000ms window (t-test, all P-values<10^{-10}).

**Computational models**

In this section, we describe the three computational learning models compared in this study. A common choice model was then used in combination with each of
these learning models to predict the probability of each subject’s choice on trial \( t \), \( c_t \), which will be presented later. All three models estimate expected outcome, \( x_t \) on trial \( t \) according to outcomes received, \( o_t \), and some free parameters. Without loss of generality, the outcome is coded as one in win trials (receiving reward or avoiding punishment) and zero in loss trials (losing reward or receiving punishment). Each of these three models formalizes learning according to a prediction-error rule:

\[
x_{t+1} = x_t + \alpha_t \delta_t,
\]

where \( x_t \) is the expected value of outcome on trial \( t \); \( \delta_t \) is the prediction error on trial \( t \) and \( \alpha_t \) is the learning rate representing the degree to which the prediction error influences the current expected outcome.

**Rescorla-Wagner model.** This model (Rescorla et al., 1972) is the simplest model among these three models, containing only one free parameter as constant learning rate, \( \rho \), bounded in the unit range, [0 1]. The Rescorla-Wagner model learns the expected outcome separately for the two actions, therefore \( x_t(c_t) \) will be updated on trial \( t \):

\[
\delta_t = c_t - x_t(c_t)
\]

\[
x_t(c_t) = x_t(c_t) + \rho \delta_t,
\]

Therefore, for this model, \( \alpha_t = \rho \) in all trials.

**Pearce-Hall model.** This model estimates a dynamic learning rate using the absolute value of prediction error (surprise). Here, we use a variant of the Pearce-Hall model proposed by Li et al. (Li et al., 2011), which replaces the constant learning rate of the Rescorla-Wagner model with Pearce-Hall associability. The resulting learning rate is:

\[
a_{t+1} = \mu |\delta_t| + (1-\mu)\alpha_t
\]

\[
a_t = \rho a_t
\]

where \( a_t \) is the associability on trial \( t \). \( \mu \) and \( \rho \) are constant parameters (bounded in the unit range) determining the step-size for updating associability and the scale of learning rate, respectively. Larger values of \( \mu \) result in faster updating of associability. It is clear that the Rescorla-Wagner model is a special case of the
Pearce-Hall model for $\mu = 0$. Given that in the Pearce-Hall model the learning rate on trial $t$ depends on the prediction error on trial $t-1$, learning rate can be seen as a proxy of the information expected to be gained before observing outcome on trial $t$ (i.e. $o_t$), having access to the prediction error on this trial, $\delta_t$. Accordingly, the regression analysis presented in the Results resembles a Pearce-Hall model when $\mu = 1$. The initial value of associability was assumed to be one.

Hierarchical Bayesian learning model. We employed the hierarchical Gaussian filter proposed by Mathys et al. (Mathys et al., 2011), which applies variational approximations to learn the hierarchical Bayesian learning model proposed by Behrens et al. (Behrens et al., 2007). A graphical illustration of this model is presented in Figure 1. This model tracks both volatility in addition to value and assumes that volatility determines the variance of value signal. More specifically, volatility has a Gaussian distribution given by

$$v_t \sim N(v_{t-1}, \eta)$$

where $\eta$ is assumed to be constant (with respect to time). The volatility modulates the probability of value, which is also given by a Gaussian distribution

$$x_t \sim N(x_{t-1}, \omega \exp(\kappa v_t))$$

where $\kappa$ and $\omega$ are two constant parameters (with respect to time). Note that instead of learning reward or punishment, this model learns probabilities in the contingency space. We assumed only one contingency trajectory, as the outcome of one action (e.g. go) always predicts the potential outcome of the other action. Subjects were aware of this dependency as they were instructed that in each trial, only one action leads to win and the other to loss. Here, without loss of generality, we assumed that the go-feedback contingency is tracked. Thus, the feedback, $f_t$ on trial $t$, is equal to $o_t$ if the choice on trial $t$ was go, and $f_t = 1 - o_t$ for trials in which a no-go response is made. Therefore, $x_t$ for this model is the expected value of the go response in the contingency space. The prediction error for learning $x_t$ is:

$$\delta_t = f_t - p_t$$

$$p_t = s(x_t)$$

in which $p_t$ is the probability predicting that go response results in a win outcome and is given by the sigmoid function, $p_t = (1 + \exp(-x_t))^{-1}$. Thus, higher value of $x_t$ lead to higher probability of go-win contingency, $p_t$. 
Mathys et al. (Mathys et al., 2011) have proved that the learning rate, \( \alpha_t \), in this model is equal to value uncertainty, \( \sigma_t \). Moreover, a nice feature of this model is that two sources of uncertainty impact the value uncertainty (and thereby learning rate) in different ways. The estimated volatility on the previous trial increases the learning rate, while noise in action-outcome contingency, \( \pi_t \), decreases the learning rate:

\[
\sigma_{t+1} = (\sigma_{t-1} + V_t)^{-1} + \pi_t
\]

in which \( \pi_t = p_t (1-p_t) \) is the variance of outcome prediction at the lower level (a binomial distributed random variable) and \( V_t \) is the volatility-related component of learning rate defined as

\[
V_t = \omega \exp(\kappa v_{t-1})
\]

where \( \omega \) and \( \kappa \) are constant parameters controlling the general scale of value variance and the extent to which value variance is affected by volatility, respectively. Therefore the learning rate on trial \( t \) is an exponential function of estimated volatility on the previous trial \( t-1 \). Since this quantity is the volatility-related part of learning rate, we have used it for all behavioral and fMRI analyses and we referred to as volatility or experienced volatility in the Results section. The volatility in this model, \( v_t \), is also assumed to be normally distributed with a mean centered at the volatility of the previous trial, \( v_{t-1} \), and a constant variance defined by another free parameter, \( \eta \).

Furthermore, Mathys et al. (2011) have shown that volatility is also updated according to a volatility prediction error signaling the log-difference between sample uncertainty and predicted uncertainty. More precisely,

\[
\varepsilon_t = \frac{\sigma_t + (x_t - x_{t-1})^2}{\sigma_{t-1} + \omega \exp(\kappa v_{t-1})} - 1
\]

where \( \varepsilon_t \) is the volatility prediction error signal on trial \( t \). It can be seen that \( \varepsilon_t \) is inversely proportional to total uncertainty on trial \( t-1 \), i.e. the sum of outcome uncertainty, \( \sigma_{t-1} \), and volatility on trial \( t-1 \), and is directly proportional to total uncertainty on trial \( t \), i.e. the sum of outcome uncertainty and the updating term \((x_t - x_{t-1})^2\). The latter represents the square of the deviation of the new sample \( x_t \) from the distribution mean, which is equal to the previous sample \( x_{t-1} \) given the Markov assumption. The volatility prediction error is more than zero when uncertainty on the current trial is more than uncertainty on the previous trial.
Here, we present those equations that are crucial for understanding and interpreting the behavioral and fMRI results of this study. For the complete list of equations, see Mathys et al. (2011). This model contains three free parameters, \( \kappa, \eta \) and \( \omega \), which are assumed to lie within the unit range (Mathys et al., 2014). We have assumed that the initial value of expected outcome and volatility is zero. The initial variance of these variables, \( \sigma_0 \), is assumed to be large (\( \sigma_0 = 10 \)). We further constrained the variance of value to lower values than the initial variance (\( \sigma_t < \sigma_0 \)). The intuition behind the latter assumption is that observing any outcome feedback could not make participant less certain about their initial estimation (before any observation) of expected value.

**Choice Model.** Each of the three learning models was combined with a choice model to generate probabilistic predictions of choice. The expected outcome values were used to calculate the probability of a go-response according to a sigmoid (softmax) function:

\[
q_t = \frac{1}{1 + \exp(-b y_t + \varphi_i)}
\]

where \( y_t = x_t \) (go) \(- x_t \) (no–go) for Rescorla-Wagner and Pearce-Hall model and \( y_t = x_t \) for the Bayesian learning model, as the latter learns probabilities in the contingency space. \( b \) is the decision noise parameter, which is a constant positive parameter encoding the extent to which learned contingencies affect choice. \( \varphi_i \) is a constant parameter representing the Pavlovian tendency to choose or avoid the go-response regardless of learned values given emotional and valence content of the cue, indexed by \( i = 1,2,3,4 \) for the four trial-types. Finally, the probability of choice, \( P_t (c_t) \), is equal to \( q_t \) if \( c_t \) is a go and it is equal to \( 1- q_t \) if \( c_t \) is a no-go response.

**Model fitting**

We fitted parameters in the infinite real-space and transformed them to obtain actual parameters fed to the models. Appropriate transform functions were used for this purpose: the sigmoid function to transform parameters bounded in the unit range (the learning parameters in all models) and the exponential function to transform the decision noise parameter in the choice model. No transformation was needed for the four constant parameters, \( \varphi_i \), of the choice model as they were not bounded.

Free parameters of each model were estimated in two stages. In the first stage, a set of parameters, \( \theta_{n, \text{MAP}}^{\theta} \), maximizing log-likelihood of data plus log-prior (maximum a posteriori, MAP) was estimated for every participant separately (\( n \) is the index of
A wide (uninformative) Gaussian prior was assumed for all parameters (with zero mean and a variance of 100). Given the large variance of this prior, the prior probability of free parameters is flat in a wide range (free parameters bounded in the unit range could vary between 0.01 and 0.99 with almost equal prior probability, and the free parameters bounded in the positive space could vary between 0.01 and 148 with almost equal prior probability). Therefore, the free parameters could be estimated without prior bias. A non-linear derivative-based optimization algorithm (as implemented in the fminunc routine in MATLAB, ©Mathwork) was used for fitting. To overcome bias of the optimization algorithm to the initial point, the optimization was repeated 40 times and the best set of parameters was selected. In the second stage, a hierarchical fitting procedure was used to fit the models to participants’ choices. An expectation-maximization algorithm was used for optimizing group- and individual- parameters in an iterative fashion, with Laplace approximation for approximating the posterior distribution (Huys et al., 2011). This method estimates the mean and the variance of parameters across all participants (group parameters) in the first step. In a subsequent step, that mean and variance is used to define a normal prior distribution of parameters and to estimate parameters of each individual participant using Laplace approximation. This procedure is then continued iteratively to reach convergence. Group parameters was initialized according to the mean and variance of the individual parameters, $\theta_{MAP}^n$, fitted in the first stage. This procedure regularizes individual fitted parameters according to group parameters, thereby decreases fitting noise and protects against outliers. The final estimated values for the group parameters, $\Theta$, were used to generate the regressors used in the fMRI analyses, as they are less biased by fitting noise. For details of the hierarchical fitting procedure, see Huys et al. (Huys et al., 2011).

**Model selection**

We employed a Bayesian model comparison approach to assess which model better captures participants’ choices. This approach selects the most parsimonious model by balancing between model fits and different levels of complexity of the models (MacKay, 2003). Notably, this procedure penalizes complexity by marginalizing over both group and individual parameters using Laplace approximation and Bayesian information criterion (BIC), respectively. Accordingly, the negative log-mode evidence (NLME) could be computed as:

$$NLME \approx - \sum_n \log P(D^n | \theta^n) - \sum_n \log N(\theta^n | \Theta, \Sigma) - \frac{1}{2} mN \log 2\pi + \frac{1}{2} \sum_n \log |H_n| - m \log(N)$$

where $D^n$ is the set of choice data for the $n$th participant, $\theta^n$ is the fitted individual parameter for $n$th participant, $\Theta$ and $\Sigma$ is the mean and variance for the group
distribution, respectively, \( m \) is number of free parameters of the model, \( N \) is the number of participants and \( \det H_n \) is the determinant of the Hessian matrix of the log-posterior function at \( \theta^n \). The log-likelihood function is the predicted probability of choice data given the model and parameters defined as \( \log P(D^n | \theta^n) = \sum \log P_t(c_t) \), where the sum is over all trials. Therefore, the first term on the right hand side of the NLME equation is how well the model predicts data. The sum of the next three terms together is the penalty due to individual parameters. The last term represents the penalty approximated for \( 2m \) (mean and variance together) group parameters using BIC. See Piray et al. (2014) for further details.

**Analysis of mean learning rate and volatility**

The means (expected values) of i) learning rate and ii) volatility component of learning rate were used for the behavioral analyses (presented in Figure 2). The mean learning rate is defined as

\[
\bar{\alpha}_i^n = \frac{1}{T} \sum_{t} \left[ \alpha_t(\theta) p(\theta | D^n) \right] d\theta
\]

where \( \bar{\alpha}_i^n \) is the mean learning rate for the \( n \)th participant and trial-type \( i = 1, 2, 3, 4 \) and \( \alpha_t(\theta) \) is the learning rate at trial \( t \) calculated by the hierarchical Bayesian learning model parameters given the parameters vector, \( \theta = [\eta, \kappa, \omega]' \). Note that the sum is over all trials belonging to trial-type \( i \) and \( T \) is the number of trials per trial-type (\( T=120 \)). \( D^n \) is the set of choice data for the \( n \)th participant. Finally, \( p(\theta | D^n) \) is the posterior after observing choice data. Similarly, the mean volatility is defined as

\[
\bar{V}_i^n = \frac{1}{T} \sum_{t \in i} \left[ V_t(\theta) p(\theta | D^n) \right] d\theta
\]

where \( V_t \) is the volatility component of learning rate. The integral in these equations could be seen as a way to weight signals computed using a set of parameters with the posterior probability that those parameters have generated data. Since these integrals are intractable, they were approximated using the importance sampling technique (Bishop, 2006). This technique is a well-known method for approximating expectations when: 1) the posterior probability is known up to a normalization constant; and 2) a sampling distribution, e.g. an approximation of the true posterior, is available. Here, we have the product of the likelihood and prior as the true, non-normalized, posterior function,

\[
F_t^n(\theta) = N(\theta) P_t(c_t)
\]
where $P_t(c_t)$ is the probability of choice on trial $t$, $c_t$, estimated by the hierarchical Bayesian learning model and $N(\theta)$ is the assumed Gaussian prior (with zero mean and diagonal isotropic variance of 100). The fitted, maximum a-posteriori Gaussian distribution serves as the sampling distribution

$$
\hat{p}^n(\theta) = N(\theta | \theta_{MAP}^n, H_{MAP}^{-1})
$$

where $H_{MAP}$ is the Hessian matrix of the log-posterior function at $\theta_{MAP}^n$. Then, the importance sampling technique approximates the intractable expectation integral as

$$
\int a_t(\theta)p(\theta | D^n) d\theta \approx \sum_l \lambda^l_n a_t(\theta^l)
$$

where $\theta^l$ represents sample $l$ drawn from the sampling distribution $\hat{p}^n(\theta)$ and $\lambda^l_n$ is the importance weight of this sample defined by

$$
\lambda^l_n \approx \frac{F_t^n(\theta^l)}{\sum_m F_t^n(\theta^m) / \hat{p}^n(\theta^m)}
$$

Therefore, the expected learning rate at trial $t$ is approximated by a weighted sum of learning rates at trial $t$ across samples drawn from the maximum a-posteriori fitted distribution. The importance weights depend on the likelihood of choice on trial $t$ as predicted by the hierarchical Bayesian model given the corresponding sample parameters.

Similarly, the integral regarding the expectation of volatility could be approximated as

$$
\int V_t(\theta)p(\theta | D^n) d\theta \approx \sum_l \lambda^l_n V_t(\theta^l)
$$

These quantities are approximated by drawing 5000 sample and used to calculate $\bar{\alpha}^n_t$ and $\bar{\nu}^n_t$. Note that the individual parameters estimated using the hierarchical fitting procedure are regularized based on behavioral data from the other subjects. Therefore it is not valid to use them as the sampling distribution (and actually for any statistical inference on behavioral effects).

**fMRI data acquisition and preprocessing**

Whole-brain imaging was performed on a 3T MR scanner (Magnetom Trio Tim; Siemens Medical Systems) equipped with a 32-channel head coil using a multi-echo GRAPPA sequence (Poser et al., 2006) [repetition time (TR): 2.32 ms, echo times (TEs, 4): 9.0/19.3/30/40 ms, 38 axial oblique slices, ascending acquisition, distance
factor: 17%, voxel size 3.3_3.3_2.5 mm, field of view (FoV): 211 mm; flip angle, 908]. At the end of the experimental session, high-resolution anatomical images were acquired using a magnetization prepared rapid gradient echo sequence (TR: 2300 ms, TE: 3.03 ms, 192 sagittal slices, voxel size 1.0_1.0_1.0 mm, FoV: 256 mm).

Given the multiecho GRAPPA MR sequence (Poser et al., 2006), the head motion parameters were estimated on the MR images with the shortest TE (9.0 ms), because these images are the least affected by BOLD signals. These motion-correction parameters, estimated using a least-squares approach with six rigid body transformation parameters (translations, rotations), were then applied to the four echo images collected for each excitation. After spatial realignment, the four echo images were combined into a single MR volume using an optimized echo weighting method (Poser et al., 2006). Noise effects in data were removed using FMRIB’s ICA-based Xnoiseifier tool (FIX), which uses independent component analysis (ICA) and classification techniques to identify noise components in data (Salimi-Khorshidi et al., 2014). Other preprocessing steps were carried out in SPM12. The T1-weighted image was spatially coregistered to the mean of the functional images. The fMRI time series were transformed and resampled at an isotropic voxel size of 2mm into the standard Montreal Neurological Institute (MNI) space using both linear and nonlinear transformation parameters as determined in a probabilistic generative model that combines image registration, tissue classification, and bias correction (i.e. unified segmentation and normalization) of the coregistered T1-weighted image (Ashburner and Friston, 2005). The normalized functional images were spatially smoothed using an isotropic 6mm full-width at half-maximum Gaussian kernel.

**Statistical analysis of imaging data**

General linear model (GLM) was used to model effects at the single-subject level (first-level analysis). We performed two GLM analyses. The first GLM looked for effects related to learning value, the second GLM looked for effects related to learning volatility.

In the first GLM, we considered four sets of four regressors each (i.e. in each set, there was one regressor for each of the four trial-types): one set was time-locked to the visual presentation of cues; one set was time-locked to the visual presentation of outcomes; one set was parametrically modulated by outcome prediction error (OPE) and time-locked to the presentation of the trial outcome; one set was parametrically modulated by experienced volatility (i.e. volatility-related component of learning rate, \( V_t \)) and time-locked to the presentation of the trial outcome. Outcome prediction error is defined as \( o_t \) in which \( y_t(c_t) = p_t \) if \( c_t \)
is a go-response and otherwise \(y_t(c_t) = 1 - p_t\). Note that \(p_t\) is the probability that the go-response is associated with a win outcome predicted by the hierarchical Bayesian model. Group parameters obtained through the hierarchical fitting procedure, \(\Theta\), were used to generate these signals. Twelve motion regressors representing six motion parameters obtained from the brain-realignment procedure and their first derivative were also included.

In the second GLM, we considered the same four sets as in the first GLM, replacing the four volatility parametric regressors with four volatility prediction-error parametric regressors, \(e_t\), time-locked to the presentation of trial outcome.

Contrasts isolating main effects of the parametric regressors and interaction effects of parametric regressors with emotion and valence were estimated at the subject-level. These contrast images were then used in a second-level GLM to make inference at the group level (t-test). Two region-of-interest analyses were performed in anatomically defined masks of the striatum and the rostral cingulate motor area. The striatum mask was created based on the probabilistic Harvard-Oxford subcortical atlas by adding accumbens, caudate and putamen masks together (thresholded at \(p<0.5\)). The rostral cingulate motor area mask was created based on a diffusion-parcellation atlas of human medial and ventral frontal cortex (thresholded at \(p<0.5\)) (Neubert et al., 2015). The ventral striatum mask used for analysis presented in the Results was created based on a connectivity-based parcellation atlas of human striatum (Piray et al., 2015).

A control analysis, including an additional regressor time-locked to the occurrence of the go-response, was implemented for each of the two subject-level GLMs. No substantial change in results was observed.

**Psycho-physiological interaction (PPI) analysis**

A PPI analysis was performed using the generalized PPI toolbox (McLaren et al., 2012). We considered twelve separate interaction terms, generated by multiplying the first eigenvariate from the seed region (dorsomedial caudate) with the three sets of four regressors each time-locked on outcome used in the subject-specific GLMs. Those interaction terms and the seed-region signal were added as separate regressors to each subject-specific GLM. Contrasts isolating interaction effects between the volatility-related PPI regressors and emotion were estimated at the subject-level, and then used for group-level inference at the second level. The dorsomedial striatal mask used in this analysis was created based on the same atlas of human striatum (Piray et al., 2015) by adding a medial cluster and a dorsal cluster identified in the caudate.
Results

Forty-four female volunteers carried out a probabilistic learning task. We only recruited women to have a relatively homogeneous sample in terms of emotional reactivity (Koch et al., 2007; Domes et al., 2010). We used data from a previously published (Li et al., 2014) probabilistic learning task focused on the association between emotional choice biases and individual differences in social avoidance. In the course of the experiment (Figure 1A), participants were presented with validated images of faces (happy or angry) and were asked to make a go- or a no-go-response (i.e. press a button, or withhold a button press, respectively) in order to obtain monetary reward or avoid monetary punishment (see Materials and Methods). Each participant was also informed about outcome valence at the start of each trial by presenting the face overlaid by a color (yellow or white) indicating whether, at the end of a trial, a win-outcome consisted of obtaining a reward or avoiding a punishment. In sum, the four trial-types differed in emotional face content (happy, angry) and in outcome valence (reward, punishment). Crucially, the response-outcome contingency was probabilistic and independent for each trial type, and reversed between blocks of 5 to 15 trials several times in the course of experiment, evoking different degrees of volatility. Within each block, the probability of a win was fixed. There were matched numbers of action-outcome contingency reversals across trial types, with 120 trials in each of the four trial types (see Materials and Methods for details). Participants learned the task effectively: performance across the group was significantly higher than chance (t(43)=2.20, p=0.03). The emotional cues did not influence overall task performance (t(43)=-0.38, p=0.71), nor participants’ bias towards go responses (t(43)=-0.40, p=0.69 – see Table 1 for additional details). However, participants processed the emotional content of those cues, as indicated by longer latencies of go-responses following the presentation of angry face cues relative to happy face cues (t(43)= 3.72, p<0.001).

Humans dissociate different types of uncertainty and track volatility

We compared three learning models to account for the observed behavioral choices and considered their ability to predict learning trajectories, namely a Rescorla-Wagner model (Rescorla et al., 1972), a Pearce-Hall model (Pearce and Hall, 1980) as implemented by Li et al. (2011), and a hierarchical Bayesian learning model (Mathys et al., 2011; Iglesias et al., 2013). Each of these three models formalizes learning according to a prediction-error rule:

\[ x_{t+1} = x_t + \alpha_t \delta_t \]
where $x_t$ is the expected outcome on trial $t$; $\delta_t$ is the outcome prediction error on trial $t$ representing the difference between experienced outcome on trial $t$ and expected outcome $x_t$; and $\alpha_t$ is the learning rate representing the degree to which the prediction error influences the current expected outcome.

The three models differ in how the learning rate is conceptualized. The Rescorla-Wager model assumes a constant learning rate throughout the experiment (Figure 1B). Accordingly, this model generates poor predictions in dynamic environments where action-outcome contingencies change (Figure 1B). In contrast, both the Pearce-Hall and the hierarchical Bayesian models update learning rate on every trial, but according to different mechanisms. The Pearce-Hall model updates learning rate according to surprise, i.e. the absolute value of the prediction error. This means that learning rate increases regardless of whether errors occurred due to contingency reversal (i.e. systematic changes in the environment) or to noise (Figure 1B). The hierarchical Bayesian model increases learning rate when a contingency reversal is detected, and otherwise decreases the learning rate gradually to protect decisions against noise (Figure 1B). In short, the hierarchical Bayesian model tracks both a model of environmental volatility, and a model of action-outcome contingency (Figure 1C, Materials and Methods). Over time, the former modulates the latter by changing learning rate. Therefore, learning rate in this model is a function of both environmental noise and environmental volatility, in which the volatility signal reflects estimated changes in the environment.

### Table 1 Logistic regression analysis of performance.

<table>
<thead>
<tr>
<th>Regression coefficients</th>
<th>Mean</th>
<th>SEM</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal response</td>
<td>0.231</td>
<td>0.105</td>
<td>0.033</td>
</tr>
<tr>
<td>Optimal response x Valence</td>
<td>-0.282</td>
<td>0.069</td>
<td>0.000</td>
</tr>
<tr>
<td>Optimal response x Emotion</td>
<td>-0.020</td>
<td>0.054</td>
<td>0.707</td>
</tr>
<tr>
<td>Optimal response x Valence x Emotion</td>
<td>0.046</td>
<td>0.047</td>
<td>0.328</td>
</tr>
<tr>
<td>Valence</td>
<td>0.038</td>
<td>0.032</td>
<td>0.247</td>
</tr>
<tr>
<td>Emotion</td>
<td>-0.010</td>
<td>0.026</td>
<td>0.693</td>
</tr>
<tr>
<td>Valence x Emotion</td>
<td>-0.002</td>
<td>0.024</td>
<td>0.941</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.704</td>
<td>0.071</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The effects of optimal response defined as a binary vector (one if the go response is the optimal response in the current block) and three non-learning factors, emotion, outcome valence and valence-by-emotion as well as their interaction on the choice are analyzed. The dependent variable encodes choices (one for go response). Therefore the coefficient corresponding to the optimal response regressor quantifies performance and its interaction with other factors quantifies modulation of performance by these factors (related to Figure 1 & 4).
Figure 1 Task timeline and computational models. A) Timeline of the probabilistic learning task. Participants had to respond (either go or no-go) after a face cue was presented. A probabilistic outcome was presented following a delay. B) The Rescorla-Wagner model, Pearce-Hall model and hierarchical Bayesian learning model conceptualize learning rate differently (upper plot). The Rescorla-Wagner model assumes that the environment is deterministic and adopts a constant learning rate. The Pearce-Hall model fails to dissociate systematic changes in the environment from noise. The hierarchical Bayesian learning model detects changes in the environment and adjusts the learning rate accordingly. The lower plots illustrate Bayesian model comparison results. The negative log-model evidence indicates how well each of the three models predicted the empirical behavioral data, controlling for each model complexity. The hierarchical Bayesian learning model outperforms the other two models (lower values indicate better fit). C) Structure of the hierarchical Bayesian learning model. This model assumes that learners adopt expectations of both value and volatility. The model also assumes that value, $x_t$, and volatility, $v_t$, have Gaussian distributions centered on their previous estimation, $x_{t-1}$ and $v_{t-1}$, respectively. Over trials, expected volatility modulates expected value by influencing its variance, which serves as the learning rate in this model. The probability of win given a go response, $p_t$, is a logistic function of expected value. The variance of value is a function of the volatility parameterized by two free parameters, $\omega$ and $\kappa$, controlling the general scale of value variance and the extent to which value variance is affected by volatility, respectively. The variance of the expected volatility is assumed to be determined by another free parameter, $\eta$. On every trial, estimation of both value and volatility are updated by an error-updating rule according to an outcome prediction error (OPE) and a volatility prediction error (VPE) signal, respectively. D) True probability sequence of win given a go response (in black) for one of the four trial-types, and the predicted probability by the Bayesian learning model. Black dots indicate actual feedback of the go response drawn from the true sequence. The lower plot illustrates volatility estimated by the Bayesian learning model. The volatility signal reflects estimated changes in the environment. It can be seen that the volatility signal is generally higher in the grey phase, when there are more changes in the response-outcome mappings. The trend in volatility reflects systematic changes in the environment, but its local changes reflects local changes in the environment. Simulations in panels b and d are based on $\eta = 0.4$, $\kappa = 0.9$ and $\omega = 0.4$. 
We compared the relative ability of the Rescorla-Wagner, Pearce-Hall, and hierarchical-Bayesian models to predict participants’ choice data in the probabilistic learning task by using hierarchical fitting (Huys et al., 2011) in conjunction with a Bayesian model comparison procedure (Piray et al., 2014) (see Materials and Methods and Table 2). For each model, this procedure calculates its evidence, a measure of goodness of fit of the model penalized by the complexity of the model (MacKay, 2003). This analysis revealed that the hierarchical Bayesian learning model outperforms the other models across participants (Figure 1B, lower plot). This finding indicates that the participants differentiated distinct types of uncertainty by tracking volatility and adjusted their learning rate accordingly.

Volatility-tracking is impaired by seeing an angry face
The previous analysis showed that participants tracked volatility. The next question is whether the presence of emotional cues influence value learning by dynamically biasing learning rate through a modulation of volatility tracking, or by introducing a tonic bias in tracking outcome prediction error, similar to the bias hypothesized to be introduced by outcome valence (reward/punishment) during learning (Daw et al., 2002; Cools, 2011). At the start of each trial, participants were emotionally cued (with happy or angry faces) and informed about outcome

Figure 1 Continued.
Emotions Induced by Seeing an Angry Face Impair Learning

Valence (reward/punishment), according to a two-by-two factorial design with four trial-types that differed in emotional face content and in outcome valence (Figure 2A). We employed two different approaches to quantify effects of emotion, outcome valence, and their interaction on learning.

First, we employed a relatively theory-neutral approach to test effects of emotion and outcome valence on learning, under minimal assumptions. Namely, we considered two learning events (on trial \(t\)) that could potentially influence choice on trial \(t+1\). The first event, a reinforcement effect, is a categorical variable indicating whether feedback corresponding to the go response on trial \(t\) was a win or loss (feedback of the two responses was inversely co-linear). The second event, an information gain effect, indicates whether new information is expected to be gained by observing the feedback on trial \(t\). This event is a categorical variable indicating whether feedback of the go response on the two trials preceding trial \(t\) was matched or not. Specifically, if the feedback on trials \(t-1\) and \(t-2\) were both loss or both win, then information gain on trial \(t\) was one, otherwise it was zero. Therefore, the reinforcement and information gain events are closely related to prediction error and learning rate in the context of reinforcement learning models, respectively. We then performed a logistic regression with choice made on trial \(t+1\) as dependent variable (coded as one for go response) and the reinforcement and information gain factors as two learning-dependent predictors. Emotion (happy/angry), outcome valence (reward/punishment), and their interaction constituted three additional learning-independent predictors. Crossing these three learning-independent effects with the two learning-related effects (reinforcement, information

| \(\eta\) | -2.621 | -0.663 | 0.890 | -1.410 | 8.149 |
| \(\kappa\) | -0.255 | 0.624 | 2.225 | 0.762 | 3.055 |
| \(\omega\) | -1.546 | -0.636 | 1.476 | -0.322 | 3.430 |
| \(b\) | -0.500 | -0.116 | 0.079 | -0.195 | 0.237 |
| \(\phi_1\) | -0.105 | 0.263 | 0.755 | 0.317 | 0.323 |
| \(\phi_2\) | -0.307 | 0.051 | 0.299 | 0.140 | 0.357 |
| \(\phi_3\) | -0.575 | -0.175 | 0.100 | -0.151 | 0.252 |
| \(\phi_4\) | -0.455 | -0.147 | 0.103 | -0.115 | 0.182 |

Table 2 Fitted parameters of the Bayesian learning model individually using maximum a posteriori (MAP) and using the hierarchical fitting procedure (HFP) (related to Figure 1).
gain) led to six additional predictors, resulting in 11 regressors in the logistic regression. See Table 3 for further details.

This analysis indicated that outcome valence modulated the reinforcement effect, whereas emotion modulated the information gain effect. Specifically, a main effect of reinforcement on trial $t$ was found on the choice made on trial $t+1$ ($t(43)=16.11$, $p<0.001$), as well as a significant interaction between reinforcement and outcome valence ($t(43)=2.57$, $p=0.014$), but no significant interaction between reinforcement and emotion ($t(43)=0.63$, $p=0.53$). The interaction between reinforcement and outcome valence occurred because the effect of reinforcement on choice was higher for rewarding trials than for punishment trials, suggesting that reward might be more easily processed as a reinforce in this task than punishment (Dayan and Huys, 2009). There was also a significant interaction between information gain and emotion on choice made on trial $t+1$ (Figure 2B, $t(43)=2.07$, $p=0.045$), with information gain being higher in happy trials than in angry trials. There was no significant main effect of outcome valence or a significant interaction between outcome valence and emotion on information gain (Figure 2B, $t(43)=1.29$, $p=0.20$ and $t(43)=0.43$, $p=0.67$, respectively; see Table 3 for additional statistics on this analysis).

### Table 3 Logistic regression analysis of the effects of learning components (Reinforcement and Information gain) on choice (related to Figure 2).

<table>
<thead>
<tr>
<th>Regression coefficients</th>
<th>Mean</th>
<th>SEM</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reinforcement</td>
<td>1.573</td>
<td>0.098</td>
<td>0.000</td>
</tr>
<tr>
<td>Reinforcement x Valence</td>
<td>-0.084</td>
<td>0.033</td>
<td>0.014</td>
</tr>
<tr>
<td>Reinforcement x Emotion</td>
<td>-0.023</td>
<td>0.037</td>
<td>0.530</td>
</tr>
<tr>
<td>Reinforcement x Valence x Emotion</td>
<td>0.058</td>
<td>0.033</td>
<td>0.083</td>
</tr>
<tr>
<td>Information gain</td>
<td>-0.015</td>
<td>0.032</td>
<td>0.637</td>
</tr>
<tr>
<td>Information gain x Valence</td>
<td>-0.050</td>
<td>0.038</td>
<td>0.204</td>
</tr>
<tr>
<td>Information gain x Emotion</td>
<td>-0.071</td>
<td>0.034</td>
<td>0.045</td>
</tr>
<tr>
<td>Information gain x Valence x Emotion</td>
<td>0.015</td>
<td>0.035</td>
<td>0.671</td>
</tr>
<tr>
<td>Valence</td>
<td>-0.107</td>
<td>0.036</td>
<td>0.005</td>
</tr>
<tr>
<td>Emotion</td>
<td>-0.005</td>
<td>0.033</td>
<td>0.874</td>
</tr>
<tr>
<td>Valence x Emotion</td>
<td>0.031</td>
<td>0.032</td>
<td>0.336</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.743</td>
<td>0.089</td>
<td>0.000</td>
</tr>
</tbody>
</table>
This analysis suggests that emotion modulates information gain. However, this theory-neutral approach only considers outcome effect and information-gain effects limited to the immediately preceding trial. This analysis also suffers from the fact that the effect of information gain defined above conflates information gain from trials on which the outcome was surprising due to the probabilistic nature of feedback, and information gain from trials on which changes in outcome signaled a contingency reversal. The hierarchical Bayesian learning model allows us to disentangle these features of information gain by separately estimating learning rate, which is a function of both estimated environmental volatility and environmental noise, and the volatility component of the learning rate. Accordingly, we considered a more refined analysis to understand how emotions influence subject-specific learning rate and volatility, as estimated within the hierarchical Bayesian learning model. We employed importance sampling (Bishop, 2006), together with an approximate subject-specific distribution estimated for parameters of the Bayesian learning model, to quantify mean learning and mean experienced volatility for each trial-type (see Materials and Methods for details).

Main effects of emotion and valence, as well as their interaction, were assessed on subject-specific mean learning rates. The effects were assessed with Wilcoxon non-parametric statistics given the non-normal distribution of mean learning rate across participants. There was a strong main effect of emotion on learning rate ($p=0.002$; Figure 2C), with no significant effect of valence ($p=0.60$) and no emotion-by-valence interaction ($p=0.67$). Participants' learning rates were significantly higher for happy-face trials than for angry-face trials (Table 4). As the learning rate is a function of both estimated environmental volatility and environmental noise, we asked whether these learning-rate biases are accompanied by a similar effect of emotion on the volatility component of learning rate. We found that this is the case: there was a main effect of emotion on experienced volatility (Figure 2D; $p=0.024$), but no effect of valence ($p=0.37$) and no interaction ($p=0.45$; Table 4). As can be seen in figure 2D, despite a significant group-level statistical reliability, there was large between-subject variability in emotional modulation of experienced volatility, possibly reflecting variance in trait factors such as trait anxiety, known to influence experienced volatility (Browning et al., 2015). Here we show that, despite those between-subjects sources of variation and despite matched volatility across trial types, participants reliably perceived the environment as less volatile during trials involving an angry face than a happy face.
Table 4  Summary statistics of experienced volatility and learning rate.

<table>
<thead>
<tr>
<th></th>
<th>R &amp; H</th>
<th>R &amp; A</th>
<th>P &amp; H</th>
<th>P &amp; A</th>
<th>Emotion</th>
<th>Valence</th>
<th>E x V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate summary statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25th percentile</td>
<td>0.743</td>
<td>0.739</td>
<td>0.742</td>
<td>0.754</td>
<td>-0.086</td>
<td>-0.045</td>
<td>-0.041</td>
</tr>
<tr>
<td>Median</td>
<td>1.226</td>
<td>1.158</td>
<td>1.178</td>
<td>1.196</td>
<td>-0.022</td>
<td>-0.001</td>
<td>0.008</td>
</tr>
<tr>
<td>75th percentile</td>
<td>2.187</td>
<td>2.159</td>
<td>2.146</td>
<td>2.072</td>
<td>0.002</td>
<td>0.024</td>
<td>0.038</td>
</tr>
<tr>
<td>Experienced volatility summary statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25th percentile</td>
<td>0.113</td>
<td>0.114</td>
<td>0.113</td>
<td>0.114</td>
<td>-0.032</td>
<td>-0.010</td>
<td>-0.012</td>
</tr>
<tr>
<td>Median</td>
<td>0.377</td>
<td>0.318</td>
<td>0.323</td>
<td>0.317</td>
<td>-0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>75th percentile</td>
<td>1.066</td>
<td>1.073</td>
<td>1.073</td>
<td>1.035</td>
<td>0.000</td>
<td>0.002</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Shown as median and quartiles as a function of valence (R: reward, P: punishment) and emotion (H: happy, A: angry). The within-subject effects of Emotion (angry minus happy) and Valence (punishment minus reward) and their interaction (E x V) are also shown (related to Figure 2).

Seeing an angry face enhances volatility-tracking in the rostral cingulate motor area, and volatility-related connectivity with the dorsomedial striatum

Following a trial outcome, a hierarchical Bayesian learning model updates estimated outcome by combining outcome prediction error (i.e. the difference between observed and predicted outcome) and a learning rate that in turns depends on current estimate of volatility. In this section, we employ regression analysis to isolate fMRI signals that correlate with these learning signals given participants’ choice, and assess the presence of modulatory effects of emotion on cerebral regions tracking prediction error and volatility.

Trial-by-trial estimates of outcome prediction error and volatility (more precisely the volatility component of learning rate, see Materials and Methods for details), were calculated for each participant based on participant’s observations and choices. The fMRI regression analysis also considered outcome prediction error at the time of outcome observation. Those two model-derived parameters (prediction error, volatility) were considered as parametric regressors, separately for each of the four trial-types, leading to 8 regressors. Eight regressors of no-interest were added to account for trial-type specific effects at the time of cue presentation (4 regressors) and of outcome presentation (4 regressors).

Volatility

Previous studies have found that the rostral cingulate motor area, an area within the anterior cingulate cortex, encodes volatility and is activated during updating
of subjective models of the environment (Behrens et al., 2007, 2008; O’Reilly et al., 2013). Therefore, we tested whether the rostral cingulate motor area, anatomically defined by using a connectivity-based parcellation atlas of medial frontal cortex (Neubert et al., 2015), is involved in volatility tracking under the current experimental conditions.
In line with previous work on volatility encoding (Behrens et al., 2007, 2008), fMRI signals in the rostral cingulate motor area correlated with volatility (Figure 3A, p<0.05, small-volume FWE-corrected for the rostral cingulate motor area mask (Neubert et al., 2015)). The main novel finding of this study qualifies how the rostral cingulate motor area tracks volatility depending on the affective state of the subject. This region shows a stronger volatility-related signal during angry trials than during happy trials (Figure 3A, p<0.05, small-volume FWE-corrected for the rostral cingulate motor area mask). Therefore, there are opposite effects of emotion on volatility-tracking in the rostral cingulate motor area and on behavior (Figure 2D). The variation across participants was also consistent with these opposing effects. Namely, greater anger-induced reduction in experienced volatility correlated with greater anger-induced increases in the cingulate motor area volatility-tracking across participants (Figure 3C). Group-wise analysis, based on median-splitting of participants over emotion effects on experienced volatility to account for the non-normal distribution of mean experienced volatility, indicated that increases in rostral cingulate motor area volatility-tracking by perception of anger was stronger in those participants experiencing a less volatile environment when exposed to an angry face (t(42)=2.16, p=0.037; Figure 3C). This finding suggests that the rostral cingulate motor area compensates for anger-induced reduction in experienced volatility by increasing the gain on the relatively noisy volatility estimation, in line with the general notion that the ACC integrates neural processing of negative affect with cognitive control (Shackman et al., 2011). Furthermore, these findings suggest that emotion effects on the cingulate motor area cannot cause the overall reduction in experienced volatility. Rather, the latter effect of emotion is likely a consequence of disrupted volatility-related computations in other brain areas.

Additional whole-brain analyses confirmed the presence of robust effects of volatility-tracking in the rostral cingulate motor area (P<0.05, FWE corrected, Table 5), as well as in the pre-supplementary motor area, and bilateral lateral prefrontal cortex, but no additional significant effects of emotion on volatility-tracking. We also performed additional region-of-interest analyses in the amygdala and on the striatum. We focused on the amygdala given its important role in emotional processing (Weiskrantz, 1956; Ledoux, 1996; Phelps and LeDoux, 2005), and previous reports on amygdala sensitivity to learning rate (Li et al., 2011). Despite the presence of clear emotion-related main effects of cue in the amygdala (bilaterally, P<0.05, small-volume FWE corrected for the amygdala mask, local maximum at -14, -8, -16), with stronger signal following presentation of the angry faces, there were no significant effects of volatility, nor interactions between emotion and volatility (p<0.001 uncorrected) in the amygdala. We also considered
the striatum, given its known role in reinforcement learning and outcome valuation (O’Doherty et al., 2003; Tobler et al., 2006; Valentin and O’Doherty, 2009), anatomically defined with an independent probabilistic atlas (Harvard-Oxford subcortical areas atlas). The dorsomedial striatum consistently tracked volatility (Figure 3D, p<0.05 FWE-corrected for the striatum mask), but there was no reliable effect of emotion on volatility in this region.

**Outcome Prediction error**
In contrast to the volatility-tracking effect found in the dorsomedial striatum, outcome prediction error was significantly correlated with the ventral striatum, extending into the dorsal putamen (Figure 3D, p<0.05, small-volume FWE corrected for the striatum mask), in line with previous work (O’Doherty et al., 2003). However, no significant effect of emotion was found on processing prediction error in the striatum (Figure 3E, p>0.001 uncorrected), suggesting that emotion does not modulate prediction error computations. A differential effect of emotion on the ventral striatum, defined using an independent connectivity-based parcellation of the human striatum (Piray et al., 2015), and on the rostral cingulate motor area is substantiated by an anatomical double-dissociation in the effects of emotion on outcome prediction error and volatility-tracking in those two brain regions. Across participants, emotion-related modulation of prediction error in the ventral striatum was significantly weaker than emotion-related modulation of volatility-tracking in the rostral cingulate motor area, as reflected in a significant interaction between region-specific computations (cingulate volatility signal versus ventral striatal prediction error signal after variance normalization) and emotion (t(43)=2.15, p=0.037). A post-hoc test revealed that this interaction is driven by significantly higher volatility signal in the rostral cingulate motor area during angry trials (t(43)=2.19, p=0.034) and no significant effect of emotion on outcome prediction error signal in the ventral striatum (t(43)=0.78, p=0.44).

Additional whole brain analyses confirmed the presence of robust effects of prediction error encoding in the ventral striatum (P<0.05, FWE-corrected, Table 5) extending to dorsal putamen, amygdala and hippocampus, as well as in the ventral orbitofrontal cortex. These prediction error effects were significantly modulated by outcome valence (p<0.05, FWE-corrected), but not by emotion (no significant voxel at p<0.001 uncorrected). Specifically, outcome prediction error signal was more strongly correlated with activity in the ventral striatum and ventral orbitofrontal cortex during reward trials than punishment trials. This finding is consistent with the suggestion that reward might be more readily processed as an instrumental reinforce than punishment (Dayan and Huys, 2009).
Figure 3 Neuroimaging findings: Emotion modulates neural correlates of volatility tracking, but not neural correlates of prediction error. A) A portion of the ACC (rostral cingulate motor area) tracked participants’ experienced volatility (in green, local maximum at -8, 24, 38), more accurately during trials with an angry face (in red, -2, 18, 38). B) Regression coefficients of experienced volatility averaged across the anatomically-defined rostral cingulate motor area. C) The emotional modulation of volatility-tracking in this region is inversely related to participants’ experienced volatility. Those participants reducing their experienced volatility when seeing angry faces showed increased volatility-tracking in the rostral cingulate motor area. D) The dorsomedial striatum (in green, local maximum at -10, 14, 2) and the ventral striatum (in blue, 16, 10, -8) tracked experienced volatility and outcome prediction error, respectively. The analysis is performed in a priori anatomical mask of the striatum. E) Regression coefficients of outcome prediction error averaged across an anatomically-defined mask of ventral striatum. The lack of emotion-related effects on the neural correlates of outcome prediction error was statistically qualified by a significant difference between the effects of emotion on the neural correlates of volatility tracking (panel b) and the effects of emotion on the neural correlates of outcome prediction error (panel e). F) The anatomical sections illustrate the rostral cingulate motor area (in red) showing increased volatility-related connectivity with a dorsomedial striatal seed region (in grey) in those trials involving angry faces (left maximum at -10, 22, 34; right maximum 10, 30, 30). In panels a, d, and f, statistical maps have been thresholded at P<0.01 uncorrected for display purposes. Error bars indicate standard error of mean. Coordinates are in MNI space.
Volatility-driven connectivity
The neural systems tracking learning rate and prediction error need to interact, eventually, as the integration of both computational elements is necessary for learning action-outcome contingencies. We tested how emotions modulate this integration. We reasoned that the dorsomedial striatum is anatomically well-placed to play this integrative role. It has dense dopamine-dependent connections with the ventral portion of the striatum that tracks prediction error (Haber et al., 2000; Piray et al., 2015), its activity correlates with volatility (Figure 3D), and it has direct connections with the rostral cingulate motor area (Draganski et al., 2008) that tracks volatility as a function of affective state (Figure 3A). We used the human striatum atlas (Piray et al., 2015) as in the above analyses to define a dorsomedial
Table 5 Whole brain analysis of main effects of experienced volatility and outcome prediction error (related to Figure 3).

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Peak (mm)</th>
<th>T(43)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced Volatility</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-14</td>
<td>4</td>
<td>68</td>
</tr>
<tr>
<td>36</td>
<td>48</td>
<td>22</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>46</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>-8</td>
</tr>
<tr>
<td>-12</td>
<td>6</td>
<td>-12</td>
</tr>
<tr>
<td>-10</td>
<td>40</td>
<td>-12</td>
</tr>
<tr>
<td>-32</td>
<td>38</td>
<td>-10</td>
</tr>
<tr>
<td>28</td>
<td>-12</td>
<td>2</td>
</tr>
<tr>
<td>-20</td>
<td>30</td>
<td>52</td>
</tr>
<tr>
<td>-54</td>
<td>-50</td>
<td>-8</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td>10</td>
</tr>
<tr>
<td>22</td>
<td>-14</td>
<td>28</td>
</tr>
<tr>
<td>20</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>-22</td>
<td>-6</td>
<td>28</td>
</tr>
<tr>
<td>18</td>
<td>-80</td>
<td>-8</td>
</tr>
<tr>
<td>-28</td>
<td>-14</td>
<td>4</td>
</tr>
<tr>
<td>-8</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>54</td>
<td>6</td>
</tr>
<tr>
<td>-26</td>
<td>-16</td>
<td>12</td>
</tr>
</tbody>
</table>

Seeing an angry face impairs volatility-learning in the right lateral prefrontal cortex
Following a trial outcome, a hierarchical Bayesian learning model updates not only its estimated action values, but also its estimation of volatility. As in the case of the update rule of estimated values, volatility updates depend on volatility
prediction errors, which in turn represent the difference between environmental uncertainty (as sampled by witnessing new outcomes) and predicted uncertainty. Given that emotion modulated experienced volatility across subjects (Figure 2D), emotion should modulate neural processes related to learning volatility. In this section, we employ regression analysis to isolate fMRI signals that correlate with volatility prediction error, and assess the presence of modulatory effects of emotion on this signal.

Cerebral responses tracking volatility prediction error, over and above tracking of outcome prediction error, were isolated with a multiple regression analysis of fMRI data. Those two model-derived parameters (volatility prediction error, outcome prediction error) were considered as parametric regressors at the time of outcome, when it is expected that predicted volatility gets updated, separately for each of the four trial-types, leading to 8 regressors. As in the previous regression model, regressors of no-interest accounting for general effects of trial types, regardless of learning signals, were also included.

Volatility prediction error correlated with activity in the right lateral prefrontal cortex (LPFC), along the inferior frontal sulcus and lateral orbital gyrus (Figure 4A, p<0.05 FWE-corrected, Table 6), corresponding to area 45 and 47 in the parcellation of Neubert et al. (Neubert et al., 2014), the connectivity patterns of which resemble macaque Brodmann areas 45A and 47, respectively. Significant effects of emotion on volatility prediction error were found in the lateral prefrontal region (Figure 4A, p<0.05 FWE-corrected, Table 6). LPFC activity related to volatility prediction error was significantly disrupted after seeing angry faces (Figure 4B). Furthermore, between-subjects variation in emotion-related modulation of volatility prediction error-related signal in the LPFC was significantly related to how experienced volatility was influenced by emotions. Individuals in whom angry cues led to larger disruption of volatility prediction error in the LPFC exhibited stronger reduction in performance (the degree to which subjects chose the optimal response, see Table 1) on trials involving angry cues (Figure 4C, t(42)=2.24, r=0.33, p=0.03). Importantly, this pattern was opposite to that observed in the rostral cingulate motor area, consistent with a putative role of this region in compensating for emotional alterations in volatility tracking. Those individuals with greater effects of angry cues on volatility-tracking signal in the rostral cingulate motor area showed weaker reduction in performance following presentation of angry faces (Figure 4D, t(42)=2.02, r=-0.30, p=0.05). This double dissociation (reflected in a three-way interaction between area (ACC versus LPFC), emotion and performance, t(41)=2.90, p=0.006) suggests that these two regions play complementary roles in integrating affective states with volatility estimation.
Figure 4 Neuroimaging findings: Emotion modulates neural correlates of volatility learning. A) A portion of the right lateral prefrontal cortex (in green, local maximum at 44, 42, -12) tracked volatility prediction error, less accurately during trials with an angry face (in red, 54, 26, 4). B) Regression coefficients of volatility prediction error (at 44, 42, -12) as a function of emotional content of cue, showing less accurate tracking of volatility prediction error after seeing angry faces. C) Emotional modulation of volatility prediction error in the lateral prefrontal cortex (44, 42, -12) is inversely related to behavioral performance across participants. D) This plot illustrates the complementary effect that emotion evokes on volatility-tracking in the rostral cingulate motor area (main effect of volatility tracking at -8, 24, 38) as a function of behavioral performance across participants. The complementary roles of LPFC and ACC in volatility processing are reflected in a significant three-way interaction between those two regions, emotion, and performance. In panel a, clusters around significant voxel (FWE corrected at P<0.05) with the uncorrected threshold of P<0.001 were displayed. Insets in panels c and d: mean and errorbars of the same data shown in two groups of participants median-splitted according to the degree of reduction in performance. Errorbars indicate standard error of mean. Coordinates are in MNI space.
Note that the region in LPFC showing significant effects of emotion on volatility prediction error signal is overlapping only slightly with the LPFC region showing significant correlation with volatility prediction error (Figure 4A). However, representation of volatility prediction error in LPFC is indeed robustly disrupted in angry trials as indicated by a significant interaction between emotion and volatility prediction error signal at the peak of the LPFC main effect (t(43)=2.34, p=0.024).

**Table 6** Whole brain analysis of main effects of volatility prediction error and its interaction by emotion (related to Figure 4).

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Peak (mm)</th>
<th>T(43)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volatility prediction error</td>
<td>44 42 -12</td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td>50 20 36</td>
<td>5.25</td>
</tr>
<tr>
<td>Volatility prediction error x Emotion (Happy&gt;Angry)</td>
<td>54 26 4</td>
<td>5.23</td>
</tr>
</tbody>
</table>

**Discussion**

This study uses a hierarchical Bayesian learning model to understand how humans learn higher order statistics of a simple environment, and how emotions influence that learning process. There are two main findings. First, emotions influence action-outcome learning by biasing estimation of environmental volatility, thereby affecting learning of action-outcome contingencies. Second, this study defines a neural circuit in which emotional processing interacts with volatility processing. Specifically, we show that an aversive emotional context compromises volatility learning in the right dorsolateral prefrontal cortex, while boosting representation of current volatility in the anterior cingulate cortex. These findings qualify the computations and the mechanisms mediating emotional influences on associative learning.

**Behavioral effects of emotion on associative learning**

We showed that formalizing human decisions requires models that consider higher order statistical regularities in the environment. Specifically, humans adjust learning rate to the estimated volatility of the environment (Behrens et al., 2007b, 2008; Nassar et al., 2010; Iglesias et al., 2013; Payzan-LeNestour et al., 2013). Importantly, we found that emotions modulate estimated learning rates and experienced volatility. Specifically, participants experience the environment less
volatile in the angry trials compared to happy trials, suggesting that emotionally aversive contexts disrupts volatility tracking.

The Bayesian hierarchical approach used in this study captures human choice behavior with a principled but relatively complex model, involving three levels of abstraction from the empirical data and three free parameters modulating these levels. This approach provides a deeper understanding of hierarchical information processing, but it relies on fitted parameters that might not be robustly estimated in each and every participant. Therefore, we also performed a simpler and relatively theory-neutral analysis using logistic regression, confirming the observation that emotions modulate information gain in participants.

**Corticostriatal circuits for updating action-outcome predictions**

Theories of associative learning suggest that emotions modulate behavioral and neural processes of learning and choice, but it remains unclear how. At the neural level, emotions could modulate structures encoding volatility or areas implementing action-outcome associations. Here, we provide empirical evidence that emotional context strongly influences volatility representation in the brain. Specifically, different components of the frontostriatal circuits implicated in learning and decision-making were found to be correlated with volatility. These components include the anterior cingulate cortex, a region previously shown to encode predicted volatility (Behrens et al., 2007b, 2008), and the striatum, a region known to implement action value learning (Delgado et al., 2000; Samejima et al., 2005; Schönberg et al., 2007). Emotional context induces differential modulations within that circuit. Emotional context modulated volatility signals in the anterior cingulate, with a stronger volatility signal during emotionally aversive contexts. We could not find evidence of this happening in the striatum. Furthermore, emotional context also modulated the connectivity between the anterior cingulate and the dorsomedial striatum, with stronger connectivity during exposure to an angry face. Given the dense connectivity of the dorsomedial striatum with portions of the striatum processing outcome prediction error, this finding fits with the notion that the dorsomedial striatum can generate trial-by-trial updates of action outcomes according to emotionally modulated learning rates and prediction errors. Future follow-up investigations will need to define a connectivity structure between this circuit and the right dorsolateral prefrontal region computing volatility prediction error.

Volatility-learning depends on volatility prediction errors, namely the difference between uncertainty (as sampled by witnessing new outcomes) and predicted uncertainty. A portion of the right dorsolateral prefrontal cortex was found to
correlate with volatility prediction errors, but only during trials involving happy faces. This interaction between emotional context and volatility-learning could be interpreted in two ways. One possibility is that the right dorsolateral prefrontal region is primarily involved in computing volatility prediction errors. Aversive emotions might disrupt these computations in this region. Alternatively, the right dorsolateral prefrontal region is primarily involved in regulating emotions on the basis of an internal model of the environment [“model-based” emotion regulation, (Etkin et al., 2015)], an instance of the known contribution of this region to model-based decision-making (Smittenaar et al., 2013). Model-based emotional regulation would be particularly taxed when the internal model of the environment is changing rapidly. The combined demands of a highly volatile environment and the presence of aversive emotions might disrupt emotional regulation in this region. Future studies manipulating a broader range of emotional contexts and volatility might distinguish between those possibilities.

**Interpretational issues**

The manipulation of outcome valence enabled us to test its influence on emotional biases of learning. For instance, some scholars have suggested separated reinforcement learning systems for appetitive and aversive learning, involving dopamine and serotonin, respectively (Daw et al., 2002; Cools et al., 2011). At the neural level, different reinforcement learning mechanisms for reward and punishment learning could result in different systems for encoding prediction error, or different systems for encoding dynamic learning rate, or both. Here, we found that outcome valence strongly modulate correlates of prediction error in the brain. However, there were neither behavioral nor neural effects of outcome valence on predicted volatility suggesting that emotional biases on predicted volatility are independent of outcome valence.

**Conclusion**

In this study, we have investigated the effects of emotion on associative learning and its influences on the neural circuitry implementing encoding and learning of changes statistical regularities in the environmental. These findings open the way to capture physiological and pathological variability during associative learning in terms of computationally well-defined parameters, and to test their neuro-biological plausibility.
Concluding remarks
Computational neuroscience provides quantitative tools for understanding brain disorders associated with deficits in learning and decision making. The mesostriatal and corticostriatal circuitry form the neural substrates of these cognitive functions. Dysfunction of elements of these circuits is associated with brain disorders such as drug and behavioral addictions, obesity, Parkinson’s disease and depression.

In this thesis, we attempted to shed light on neural circuitry implementing computations underlying reward learning and choice. We employed pharmacological manipulation, cognitive tasks and neuroimaging techniques together with state of the art computational models of brain and behavior.

**Summary**

Here, I summarize the main findings:

In chapter 2, we focused on functional architecture of the human striatum and its modulation by dopamine. We found a hierarchical organization along the ventrodorsal axis modulated by dopamine. Furthermore, trait impulsivity was associated with a specific component of this circuitry, namely the dopaminergic modulation of the input from the ventral to the dorsomedial striatum.

In chapter 3, we studied Pavlovian and instrumental learning in Parkinson’s disease patients diagnosed with impulse control disorders triggered by dopaminergic medications. Using a reinforcement learning actor-critic model of the striatum (Barto, 1995), we found that impulse control disorders are associated with model parameters related to stimulus valuation (Pavlovian learning). Specifically, Parkinson’s disease patients with impulse control disorders exhibited lower learning rate from negative prediction errors in the critic.

In chapter 4, we focused on neuroanatomical traits corresponding to individual differences in exhibiting goal-directed and habitual forms of decision making quantified using model-based and model-free reinforcement learning accounts (Daw et al., 2011). Using diffusion-based structural connectivity, we found that individual differences in the degree of model-based control are predicted by neuroanatomical projections from the ventromedial prefrontal cortex to the medial striatum. Individuals with stronger ventromedial prefrontal afferences to the medial striatum are more likely to employ a model-based strategy in decision making.
In chapter 5, we focused on emotional modulation of learning in volatile environments. We employed a hierarchical Bayesian learning model (Behrens et al., 2007b; Mathys et al., 2011) and found that emotions modulate subjective experiences of environmental volatility. Specifically, experienced volatility was lower in emotionally aversive contexts resulting in lower learning rates in these contexts. Negative emotions suppressed signals vital for volatility tracking in the dorsolateral prefrontal cortex and biased the neural circuitry implementing action-outcome contingency learning by modulating the volatility encoding in the anterior cingulate region and its volatility-dependent interaction with the dorsomedial striatum.

Contributions to computational psychiatry

In the first chapter, I noted three levels within the computational psychiatry program to identify pathological computational processes in patients with disorders related to learning and choice, such as addiction and impulse control disorders. These levels are concerned with 1) computational modeling of healthy brain; 2) mapping potential endophenotypes of diseases into priors (or parameters) of the computational models; 3) modeling brain disorders as the pathological state of maladaptive interactions between priors and risky environment in which priors are suboptimal.

A long-term perspective that this thesis is based upon was to shed light on candidate cognitive and neural mechanisms and associated computations that might predispose individuals to compulsive disorders. Therefore, in this thesis, I focused on computational modeling of some neurocognitive traits that are among candidates of addiction endophenotypes. In chapters 2, 3 and 4, we aimed to identify priors biasing computations implemented within the corticostriatal circuitry that are related to these endophenotypes (Dalley et al., 2011; Robbins et al., 2012). Specifically, in chapter 2, we focused on the striatal architecture, its dopaminergic modulations and its interaction with impulsivity. In chapter 3, we studied maladaptive computational processes and their modulations with dopamine in Parkinson’s disease patients with impulse control disorders. In chapter 4, we studied components of corticostriatal circuitry predicting individual differences in exerting model-based and model-free control on actions.

In chapter 5, we studied computational modeling of emotional biases of learning and choice. Here, we aimed to fill the gap in computational modeling of emotion. Therefore, this chapter lies within the level one. It is necessary to fill this gap as individual differences in emotional traits might be related to different aspects of drug addiction, for example initiation of drug use and relapse (Khantzian, 1997).
Putting together, findings of this thesis shed light on candidate computational processes in neural circuitry implementing learning and choice in patients with impulsive/compulsive disorders.

**Future directions**

In this section, I wish to mention some future directions arisen by this thesis with a focus on computational modeling research on psychiatric disorders.

In chapter 2, we found that a specific element of mesostriatal circuitry predicts individual differences in trait impulsivity in healthy volunteers. Specifically, we found that the degree of D2-dependent dopaminergic modulation of ventral to dorsomedial striatal connectivity predicts trait impulsivity. Given previous studies showing a link between impulsivity and propensity to compulsive drug abuse (Everitt et al., 2008), our results suggest that high sensitivity of coupling between ventral striatum and dorsomedial striatum (caudate nucleus) to dopaminergic challenges might be a predisposing neurobiological trait of addiction. This is particularly interesting given recent evidence that the dorsomedial striatum plays a critical role in drug seeking behavior during early phases of drug exposure, when behavior is still goal-directed and sensitive to outcome devaluation. For example, Corbit et al. (2012) have shown that in early phases of alcohol exposure, drug seeking behavior in rats is dependent on dorsomedial-, but not dorsolateral-, striatum. In humans, consistent with these findings, an aberrant intrinsic connectivity between ventral striatum and caudate in abstinent heroin users has been found (Xie et al., 2014), although it is not clear from this study that this is a predisposing neurobiological trait or only a consequence of drug exposure. Our findings suggest that one possible mechanism through which drugs could modulate goal-directed drug taking actions during early phases of drug intake in highly impulsive individuals is the dopamine-dependent modulation of the dorsomedial striatum by the ventral striatum. Future works should address this question in individuals at risk of impulsive and compulsive disorders.

In chapter 3, we found that Parkinson’s disease patients with impulse control disorders exhibit deficits in learning states values from negative prediction errors. This finding is generally consistent with the notion of compulsivity, which is the pathological state of ignoring harmful consequences of a strong habitual response. However, consistent with recent theories emphasizing the interactions between Pavlovian and instrumental learning in developing addiction, here we found that deficit in these patients is specific to Pavlovian learning of state values. As only a
small subset of Parkinson’s disease patients medicated with dopaminergic medications develop impulse control disorders (Dagher and Robbins, 2009), it is important to identify risk factors in these individuals making them vulnerable to dopaminergic medication. Given our results, a question for future studies is whether this specific deficit in Pavlovian learning is also present in these patients before onset of Parkinson’s and dopaminergic medication, and whether dopaminergic medication exacerbate this deficit.

There is an increasing interest to describe brain disorders as a pathological state of the balance between goal-directed and habitual instrumental control. Compulsive disorders such as drug addiction, binge eating and obsessive-compulsive disorders are conceptualized as a progressive pathological transition from goal-directed action control to habitual stimulus-driven habits (Everitt et al., 2008; Smith and Robbins, 2013; Gillan and Robbins, 2014). Some manifestations of Parkinson’s disease and Tourette syndrome have also been hypothesized to be due to the imbalance between goal-directed and habitual control (Redgrave et al., 2010; Jahanshahi et al., 2015). Therefore, computational modeling of these modes of action control as different reinforcement learning systems is a step forward for quantifying the balance between the two systems (Voon et al., 2014). However, one important question is how, neurobiologically and computationally, the balance goes awry in these disorders. Importantly, the main focus of original works on computational modeling of goal-directed and habitual control was on normative computational mechanisms underlying the balance between the two systems (Daw et al., 2005; Keramati et al., 2011). Future works should address the pathological computations causing the imbalance between the two systems in patients with these brain disorders and its neurobiological correlates. Our finding indicate that the top-down connectivity from the ventromedial prefrontal cortex to the medial striatum through cingulum white matter might be the neurobiological substrate of passing information necessary for these computations. This possibility should be also addressed in future research.

Modern research in cognitive neuroscience of emotion describes it as a modulating and interacting system with learning and decision making (Phelps et al., 2014; Etkin et al., 2015). A wide range of psychiatric disorders also manifests as simultaneous pathological state of emotional processing, learning and choice. In chapter 5, we showed that emotions modulate the neural circuitry implementing learning uncertainties and causal structures in the environment. An important question is what elements of this circuit are particularly related to emotional traits, such as anxiety and social avoidance, which might predispose psychiatric disorders such as addiction and depression (Belin et al., 2015). In addition, future
works should investigate the role of different neuromodulatory systems, such as noradrenergic and acetylcholine (Yu and Dayan, 2005), in learning uncertainties and their interaction with emotional processing within this circuit.

In this thesis, we emphasized on mappings of cognitive and neuronal traits into parameters and priors of computational models, which could help to construct “computational traits” with potential diagnostic value. This perspective heavily depends not only on our ability to define such models, but also on statistical methods making inference about validity and robustness of estimated parameters across individuals and patient groups, as well as making inference on generalizability of computational models. Currently, there is a lack of such robust statistical tools. Future works in computational psychiatry should address this issue by providing appropriate tools for the academic psychiatry community.
Bibiliograpgy
Bibliography


Bibiliography


164


Nederlandse Samenvatting
Acknowledgement
About the author
List of publications
Donders Graduate School for Cognitive Neuroscience
Nederlandse Samenvatting

De computationele neurowetenschap biedt kwantitatieve instrumenten waarmee men hersenstoornissen kan bestuderen die gepaard gaan met aandoeningen van cognitieve functies zoals leren en besluitvorming. Deze cognitieve functies worden ondersteund door de meso-striatale and cortico-striatale netwerken in de hersenen. Dysfunctie van onderdelen van deze netwerken worden geassocieerd met hersenaandoeningen, zoals verslaving, obesitas, de ziekte van Parkinson, en depressie. In dit proefschrift hebben we getracht om de hersen-netwerken in kaart te brengen die de berekeningen uitvoeren die ten grondslag liggen aan beloning-gestuurd leren en besluitvorming. We hebben hiertoe een verscheidenheid aan technieken ingezet, waaronder farmacologische manipulatie, cognitieve taken, en beeldvormende technieken, in combinatie met computermodellen van hersenfunctie en gedrag. Hieronder geef ik een samenvatting van de belangrijkste bevindingen:

In hoofdstuk 2 hebben we ons gericht op de functionele organisatie van het striatum en hoe deze wordt beïnvloed door dopamine. Met zogenaamde resting state scans hebben onderzocht hoe verschillende onderdelen van het striatum met elkaar in verbinding staan. Onze resultaten ondersteunden een hiërarchische organisatie van het striatum, waarbij informatie wordt uitgewisseld tussen ventrale en dorsale striatum, via het dorsomediale striatum. Met farmacologische manipulaties konden we aantonen dat dopamine deze verbindingen moduleert. Daarnaast vonden we dat de karaktereigenschap impulsiviteit samenhangt met een specifiek onderdeel van dit circuit, namelijk de dopaminerge invloed op de input van het ventrale naar het dorsomediale striatum. Deze resultaten helpen ons begrijpen waarom hoog impulsieve mensen gevoeliger zijn voor verslaving.

In hoofdstuk 3 hebben we Pavloviaans en instrumenteel leren onderzocht bij Parkinson patiënten die gediagnosticeerd waren met impuls controle stoornissen geïnduceerd door dopaminerge medicatie. Door middel van een reinforcement “actor-critic” leer model van het striatum, toonden we aan dat impuls controle stoornissen geassocieerd zijn met parameters van het model welke gerelateerd zijn aan stimulus waardering (Pavloviaans leren). Om precies te zijn, Parkinson patiënten met impuls controle stoornissen lieten een lager leertempo zien bij foutieve negatieve voorspellingen in de “critic”.

In hoofdstuk 4 hebben we onderzocht welke neuro-anatomische kenmerken correleren met de verschillende manieren waarop proefpersonen keuzes maken. We maakten gebruik van diffusion-based structurele connectiviteit en vonden dat de neuroanatomische projecties van de ventromediale prefrontale cortex en het
mediale striatum voorspellen in welke mate proefpersonen gebruik maken van model-based controle. Bij individuen met sterkere verbindingen van de ventromediale prefrontale cortex naar het mediale striatum is het waarschijnlijker dat ze gebruik maken van een model-based strategie bij het maken van keuzes.

In hoofdstuk 5 hebben we onderzocht hoe aversieve versus appetitieve emoties het leren beïnvloedt in een onzekere context. We gebruikten een zogenaamde hierarchische Bayesian leermodel als computationele methode en vonden dat emoties de subjectieve ervaring van contextuele volatiliteit beïnvloedt. Om precies te zijn, we vonden dat de ervaren volatiliteit lager was in een aversieve context, waardoor de leersnelheid in deze context ook lager was. Dit effect hing samen met de neurale bevindingen. Aversieve emoties onderdrukten de signalen in de rechter laterale prefrontale cortex, die cruciaal zijn voor het leren over volatiliteit. Dit beïnvloedde het neurale circuit dat een rol speelt bij het leren van doelgericht gedrag door het coderen van volatiliteit in de anterieure cingulate cortex en de interactie hiervan met de dorsaal-mediale striatum te moduleren.
Acknowledgement

It is always difficult to wrap up a thesis. It is even more difficult to find suitable words to thank people who helped you in this journey.

Roshan and Ivan, it was an amazing journey together throughout my PhD. I am extremely thankful to you for giving me the opportunity to grow freely but providing me with your advice and expertise when it was needed. I am also grateful to you for tolerating my scientific stubbornness.

Ivan, “il padrino”, you have been a true scientific mentor and one of my best friends in donders. Your contribution to this thesis is evident in almost every page. Your deep and true care about your students, even those like me that drop by every day, is invaluable. Your style of supervision somehow maximizes creativity and critical thinking in your students. These are what make you a true mentor.

Roshan, you have always been a source of inspiration for me, even before starting my PhD. Your deep knowledge about key elements of this work had a huge impact on the focus and outcome of this thesis. I learned a great deal from you not only about brain and cognition, but also about how to communicate science and how to manage time. I am grateful for your continued support and for being a constant source of encouragement.

It is always fun to be in the Motivation and Cognitive Control lab. I would like to thank all of you for making this group excellent and fun. Guillaume, thanks for managing the labmeetings amazingly. You are a great leader and a modest and fantastic scientist. Lieneke and Monja, thanks for all your contribution to the social atmosphere of the lab. This group can never be so fun without you. Hanneke, Romain and Bram, thanks for your insights and feedbacks. Guillaume, Dirk, English Jen, Dutch Jen, Monique, Joost, Esther and Verena, your research has always been stimulating and inspiring. Thanks for being so great!

I would like to thank members of Ivan’s research group, especially Lennart, Inge, Sasha, Arjen, Anke Marit and Loek. I have always enjoyed chatting with you and listening to your constructive comments. I would like to express my appereciation to Lennart for sharing his wide knowledge in statistical methods in neuroimaging during groupmeetings and FAMs.

Verena, Marieke and Hanneke, this thesis was not possible without our fruitful collaborations. I am very grateful to you.
Sophie, Josi, Andrea and Joyce, thanks for your great help in our social learning project, which unfortunately I did not manage to publish in this book. Sophie, you have been a great and responsible student researcher.

Donders is a fantastic place to work. I was delighted to be able to meet some amazing people almost every day. I would like to specially thank all those fantastic people you can easily find in the canteen or FADs: Eric, Anke Marit, Daan, Ruud, Sebo, Jeroen, Marieke, Ian and Tom.

Donders is also an absolutely first class environment for doing research. This is partly because researchers in DCCN are supported by a great team of technical and administrative staff. Thanks for your support. I would like to thank Tildie for her care about everyone in DCCN. I also like to thank Paul for amazing help and support in the basement of scanners.

Guillaume, Yoolla, Monja and Lieneke, I am grateful to you for your help in organizing the computational psychiatry symposium and my defense. Guillaume and Yoolla, it it a pleasure to have you as paranymphs. Anke Marit, Monique, Bram, Verena and Marieke, thanks for your help with the “Nederlandse Samenvatting” of this book.

Yoolla, thank you for being a great friend during my years in Nijmegen. It has always been fun hanging out with you and Elnaz. Of course, I enjoyed every second of beating you in our game nights!

I wish also to thank Amir and Yashar, for fruitful discussions on many key substances of this thesis before and during my PhD.

I would also like to thank my family. Mom, dad, Ghasem and Parastoo, thanks for your great support during all these years. Parastoo, also thanks a lot for taking care of many bureaucratic issues in Tehran involved with my life abroad. Sajad, you have been a true friend that every person wishes to have. Thanks for the nice cover you designed for this book.

Above all, Samira, I am not really sure how I can thank you. You are a smart and fun person, a fantastic friend and a supportive and caring partner. Your extreme support throughout all these years has been invaluable. I am deeply thankful to you. You are a unique and endless source of true happiness, strength and love for me.
About the author

Payam Piray was born in Ray, Iran, on May 7th, 1985. He studied electrical engineering in University of Tehran, where he got a bachelor degree with a focus on control engineering in 2008. He then pursued a research master degree in control engineering with a focus on intelligent systems and defended his thesis in 2010 (with Honors). During this time, he studied computational modeling of addiction. In 2011, after one year of studying neuroscience in University of Southern California (with a USC neuroscience Fellowship), he moved to Nijmegen and started his doctoral research in cognitive neuroscience at the Donders Institute. During this period, he investigated neural mechanisms of learning and decision making under supervision of Roshan Cools and Ivan Toni, of which this dissertation is the result. Since December 2015, he is a post-doctoral researcher at Donders Institute, where he studies individual differences in cognition and its neural substrate.
**List of publications**


Donders Graduate School for Cognitive Neuroscience

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc.. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit: http://www.ru.nl/donders/graduate-school/donders-graduate/