Helga Anne-Marike Schiffer-Maraun

# Learning from the Unexpected: Neural Signatures of Perceptual Prediction Errors in the Cortico-Basal Ganglia-Thalamo-Cortical Loops

2012

Fach: Psychologie

Dissertationsthema: Learning from the Unexpected: Neural Signatures of Perceptual Prediction Errors in the Cortico-Basal Ganglia-Thalamo-Cortical Loops

Inaugural-Dissertation
Zur Erlangung des Doktorgrades
im Fachbereich Psychologie und Sportwissenschaften
der Westfälischen Wilhelms-Universität Münster

Vorgelegt von
Helga Anne-Marike Schiffer-Maraun, geb. Schiffer
aus Koeln
-2012-

Dekan: Prof. Dr. Markus Lappe

Erste Gutachterin: Prof. Dr. Ricarda Schubotz

Zweiter Gutachter: Prof. Dr. Markus Lappe

Tag der muendlichen Pruefung: ________________________________________

Tag der Promotion: ________________________________________

## Acknowledgments

First of all, I wish to thank my mentor Ricarda Schubotz for sheer endless support and trust, and for being an inspiration in so many ways. I further wish to express my sincere gratitude to Professor von Cramon for enabling my research scholarship at the Max Planck Institute, for his guidance and decisive help. I cordially thank Professor Markus Lappe for his interest in my work and the marking of this thesis. I further thank Moritz Wurm, Katja Kornysheva and Christiane Ahlheim for enlightening discussions, their invaluable friendship and for bringing the funk into work, and back to Cologne. I wish to thank Philipp Schiffer for hilarious and uplifting remarks on work as a scientist. I wish to thank Anne Kuehn and Erik Noormann for helping me getting to grips with scripting. I thank those people who contributed enormously to the coming into existence of my projects, Elena Hoefeler, Melike Bayraktar, Kirstin Ulrichs, Kim Krause, Janji Yokeeshwaran, Luisa Donner and Alexander Wagner, with special regards to Kim Krause and Kirstin Ulrichs, who worked harder than any reasonable person could expect. I thank my parents, who are great. Lastly, I thank Thomas Maraun, for private reasons.

**0 Abstract** **7**

**1 Introduction** **8**
1.1 The Myth of the Given 8
1.2 Early Forward Models 10
1.3 The Brain as an Inferential Bayesian Machine 14
1.3.1 The Influence of Information Theory 16
1.3.2 Hypotheses on the 'True State of the World' 17
1.3.3 Bayesian Model Adaptation: Surprise and Shannon Entropy 19
1.4 Predictive Coding 21
1.4.1 Predictive Coding in Action Observation 25
1.4.2 The Predictive Coding Explanation for Evoked Brain Activity 27
1.5 Reward Related Prediction Errors 28
1.5.1 Surprise Incites Learning – the Rescorla-Wagner Model 28
1.6 Prediction Error Driven Learning: the TD-algorithm 29
1.7 Dopamine 31
1.8 The Basal Ganglia 32
1.8.1 Cortico-Basal Ganglia-Thalamo-Cortical Loops 34
1.8.2 Basal Ganglia Pathways 37
1.9 Summary and Research Questions 39

**2 Research Articles** **41**
2.1 Caudate nucleus signals for breaches of expectation in a movement observation paradigm. 41
2.2 Neural Changes When Actions Change: Adapatation of Strong and Weak Expectations 54
2.3 Surprised at all the Entropy: Hippocampal, Caudate and Midbrain Contributions to Learning from Prediction Errors 70

**3 Discussion** **103**
3.1 Prediction Errors in the Basal Ganglia 103
3.2 Internal Forward Model Projections in the Basal Ganglia 104
3.2.1 Neuroanatomical Considerations Concerning Internal Forward Models 105
3.2.2 Decision Making Theory and Internal Forward Models in the Striatum 106
3.2.3 The Supplementary Motor Area: Internally Triggered Forward Models 107
3.2.4 Lateral Premotor Cortex Predictions 109
3.2.5 Clinical Studies 110
3.2.6 A proposal of a Clinical Study to Test Implications 112
3.3 Midway summary 113
3.4 Learning from (Prediction) Errors 114
3.4.1 The Relationship between Errors and Prediction Errors 115
3.4.2 Error-related Research 116
3.4.3 The ACC in Error Research 116
3.4.4 Neuroanatomy and Neurotransmitters of Errors 118
3.4.5 Proposed Study to Dissociate Errors from Prediction Errors 120
3.5 Applying Computational Models 122
3.5.1 Interpretative Concerns 122
3.5.2 Model Competition 125
3.6 In Psychological Terms 126
3.7 Final Remarks 128

**4 Appendix** **129**
4.1 The Kullback-Leibler Divergence 129
4.2 The TD-algorithm 130

4.3 Glossary 132
4.4 Abbreviations 135

**5 References 136**

**6 List of Figures: 145**

**Curriculum Vitae 146**

**Declaration 147**

## 0 Abstract

Perception and learning are two topics of psychological and neuroscientific research that have long been considered separately. Current approaches that focus on the reliance of perception as well as learning on internal forward models bridge the gap between these two avenues of research. Internal forward models are considered to underlie perceptive processes and predictions errors that result when these internal models are not in accordance with the sensory input serve the adjustment of perception; adjustment of internal forward models according to prediction errors ultimately results in learning. From a neuroanatomic point of view, activity in the basal ganglia has been shown to be responsive to prediction errors in reward-related forward models, but rarely been implicated in not-reward related prediction errors.

The present thesis contains experiments that investigated whether prediction errors in perception are also signalled in the basal ganglia. Further, we tested, what factors determine learning from perceptual prediction errors. The latter issue was investigated with regard to the both, the solidity of internal forward models and the reliability of internal model violating information. The results indicate that activity in the basal ganglia signals for prediction errors in perceptual paradigms. The amount of information in favour of the internal forward model seemed to influence learning from prediction errors. The validity of the prediction error eliciting information was shown to influence the adaptation of the internal model.

The results are discussed with regard to the anatomic structure of the basal ganglia. The idea that emerges from these results is that weighted internal forward models of external actions may be generated in the basal ganglia and that this generation is possibly modulated by the dopaminergic innervation of the structure.

# 1 Introduction

## 1.1 The Myth of the Given

Psychological accounts of brain function have often encompassed a triad of brain-modules, namely peripheral perception, action, and central cognition. But this modular view has been challenged (see Hurley, 2001 for a review). As Hurley points out, theories on action have often neglected perception, while research into perception has not been concerned with action. Behaviorism for example has shied from discussing perception, as it states that internal processes cannot and need not be inferred (Watson, 1913). Information theory on the contrary described perception as a process that minimized the uncertainty about the environment (Garner, 1975).

The influential account of perception delivered by Gibson regards perception as a means to determine the affordances of our environment and action as a means to enhance perception (Gibson, 1986). Thus, we find an instrumental interaction between these processes. This Gibsonian view includes an influential notion about perception: perception is not necessarily correct, i.e., perception does not encompass everything that exists. And: perception depends on the observer. This statement is made explicit when regarding what we (as humans) can perceive, and what we cannot. Ultraviolet light is not visible to the human observer, but bees and bumblebees use it. These observer-dependent affordances are central to Gibsonian theories (Gibson, 1986), and we see in the example of bees, that they also elicit motor behavior. This functional account of perception suggests that perception is (putatively also evolutionary) fitted to detect meaningful events for action. However, even though this account stresses instrumental interdependence of action and perception, it still regards them as two modules (Hurley, 2001, 2006).

Parenthetically, even though functional perception enables functional actions, not all kinds of perception aim at a concrete action (Schubotz, 2007). Depending on circumstances, humans engage in “watching and listening”, for example watching the clouds drift by, the leaves falling, listening to music or birdsong, and rustling winds.

Moreover, even within one observer, it is a long-standing notion that perception is subject to learning (cf. Berkeley, 1709; Helmholtz, 1866). A classic psychological experiment concerning this idea investigated the adaptation and after effect that result from wearing goggles that displace vision. Wearing these goggles usually leads to a compromised ability to perform visually guided aim or reach movements. However, this effect disappears after a while. This adaptation can be understood as perceptual learning (Kornheiser, 1976). Yet, a correspondence of visual input and motor activity is present in these experiments. Optical illusions seem to play a minor part in everyday life and are far more prominent in experimental settings, implementing two-dimensional displays (cf. Guski, 1996). This raises the question: How do we learn to perceive functionally?

Interestingly, an account of both, the proximity of action and perception, even in terms of neural coding, and the supposed account of learning comes from motor control theory (Ghahramani, Wolpert, & Michale, 1997; Wolpert, Diedrichsen, & Flanagan, 2011; Wolpert, Ghahramani, & Jordan, 1995). Even very early research into motor functions, conducted for example by von Helmholtz, Purkine and von Uexkuell stressed how the perception of movement is dependent on motor acts (Gruesser, 1986). I will give a short account of these findings and their development in the next chapter. Afterwards, I will explain that action and perception are now both understood to rely on predictive processes; processes that can even be located in the same neural circuits (Hurley, 2006; Schubotz, 2007). The current thesis is dedicated to the question of how perceptual predictions related to motor acts are revised when they fail to deliver correct predictions.

## 1.2 Early Forward Models

Recent years in neuroscientific research have seen a tremendous upsurge in the use of predictive models of brain functioning (Friston, 2010; Huang & Rao, 2011; Rao & Ballard, 1999; Summerfield et al., 2006a; Summerfield, Trittschuh, Monti, Mesulam, & Egner, 2008), but the notion of the brain as a predictive machine is not new. It rests on the concept of internal forward models. The idea of forward models can be found in the form of *efference copies*, discussed more than 60 years ago (Sperry, 1950; Von Holst & Mittelstaedt, 1950). Von Holst and Mittelstaedt argued that every motor command results in an efference copy that predicts the afference, or reafference, which will result from the execution of the motor command. The authors described for example how efference copies ensure that eye movements do not result in the impression of a moving world. If a central organ, later renamed controller, initiates an eye movement, an efference copy accompanies the motor command and predicts how the representation on the retina (the reafference) will change as a result of the movement. If this change occurs, the efference copy and reafference cancel each other out, resulting in the impression of a stable world. If, however, the bulbus of the eye was moved externally, for example by using forceps, the lack of a motor command from a controller means that no efference copy is issued, resulting in an unfiltered reafference. Hence, the impression of a moving world is generated, although it is the eyeball that has been moved.

In its current form motor control theory still posits that an *internal forward model*, sometimes called emulator (Grush, 2004), predicts the internal state that is associated with a motor command (Ghahramani, et al., 1997; Haruno, Wolpert, & Kawato, 2001; Wolpert, et al., 1995; Wolpert & Miall, 1996). In motor control, each forward model is paired with an inverse model, which associates a (desired) state with the according motor command to bring about this state (Haruno et al., 2001; Wolpert & Kawato, 1998). A third type of model, the forward sensory model, specifies the expected sensory feedback from this predicted state (Wolpert & Ghahramani, 2000)[1]. Importantly, these internal models can be run off-line, that is without an actual motor command being issued (Wolpert & Miall, 1996). This provides the ability to predict future body states and its constituent sensory consequences. If the modelled future sensory consequences that represent the 'state' diverge from the desired future state (and the associated sensory consequences) this yields the anticipation of a prediction error. These anticipated prediction errors of future states can be used to adjust motor commands (Wolpert & Miall, 1996). The inverse models can be used to imitate behaviour, because they allow the matching of a perceived state of an actor to the according motor command within the observer's motor system (Wolpert et al., 2003). To that end, the visual input must be mapped to a state that concurs to this input. The associated inverse model can then represent the necessary motor command. This also posits an explanation why action observation causes activity in the observer's motor system (Jacob & Jeannerod, 2005; Jeannerod, 1995; Miall, 2003; Wolpert et al., 2003). The underlying idea is that the perception of the consequences of a movement leads to an inferential process that determines the inverse model that would be needed to achieve the perceived transformation (Wolpert et al., 2003). Because forward models are supposed to follow a hierarchical structure, with the higher hierarchy levels coding for action

[1] We will later see that this three-fold distinction has been challenged in later accounts of motor control (Friston, 2011)

goals, the activation of inverse models during forward model estimation was proposed to a activate a neural representation of the goal of the perceived action (Miall, 2003; Wolpert et al., 2003). Therefore, it was reasoned that the activity in the motor system that accompanies action observation is related to the recognition ('interpretation' according to Csibra, 2007) of action goals. This argument relies to a certain degree on some evidence that the coding in the premotor cortex of the observer is in a way predictive, here in the sense of forecasting. The fact that the activity in the motor system during action observation resembles that during action execution, and the theory that the motor system runs internal forward model during action execution, does not necessarily warrant the interpretation that the action observation incites an internal *forward* model of the observed action. The neural activity in the motor system could mirror the perceived stage, but not contain predictions on the next action step. However, a number of studies in the macaque and in humans support the theory of a forward account of action observation. An influential review (Keysers and Perret, 2004) notes activity in the monkey brain area F5 that supports the theory of forward modelling in the motor system during action observation. Area F5 is supposed to be homologue to a portion of the ventral human lateral premotor cortex (Picard & Strick, 2001). Keysers and Perrett, (2004) described how a neuron in area F5 fired in response to observing a human hand reaching behind a screen, only if the monkey had previously observed how an object had been placed behind the screen. The activity in the neuron is associated with the manipulation of objects and the perception of someone else manipulating the object. Accordingly, the reaching for the hidden object could elicit the neuronal firing because of the associated manipulation of the object. Importantly, this response is absent when the monkey has not observed how an object is placed behind the screen. We thus find evidence that the reaching for the assumed location of an object incites the internal model of the manipulation of the given hidden object. Different evidence for forward modelling of observed actions itself comes from a study by Flanagan and Johansson (2003). The authors could show that the observer of an action makes the same, i.e. predictive, saccades as the actor, as opposed to movement following 'reactive' saccades. Similarly, Kilner and colleagues (Kilner, Vargas, Duval, Blakemore, & Sirigu, 2004) could show that readiness potentials to observed movements increase before a predicted movement came into effect. Taken together these findings indicate that the internal forward models do not only predict the sensory, or sensorimotor, consequences of our own actions, but also the sensorimotor consequences of perceived actions.

Motor control theory has been applied successfully to a large number of motor related research questions (see Wolpert et al., 2011, for a recent review), but also been criticized. The one counter argument targets motor control theory's discrimination between forward models, sensory forward models and inverse models (Friston, 2011): these could be replaced by one forward model and its Bayesian inversion (Friston, 2011), as in the perceptual paradigms that I will discuss in the chapter *Hypotheses on the 'True State of the World' (1.3.2)*. The second critique concerns the difficulty of motor control theory to explain how visual input concerning another person's actions translates to the hidden states that the proposed inverse model uses (Friston, 2011). Motor control theory, while apt to explain many phenomena of motor control is naturally limited to the motor system. In clear terms, this model is not designed to explain perception per se. Moreover, the fact that the model is rooted in the motor system has incited critique concerning its application to social inferences, i.e. inferences concerning intentions of observed actors that extend further than the goal of an action. This function is not supposed to be coded for in the motor system (Jacob & Jeannerod, 2005). In the next chapter, I will present an alternative account of action perception that rests on the predictive coding account. Predictive coding relies on a Bayesian inference account of perception. I will therefore introduce the Bayes theorem first, before progressing to predictive coding and predictive coding in action perception.

## 1.3 The Brain as an Inferential Bayesian Machine

Originally, the Bayes theorem was applied to hypothesis testing. The starting point to understanding the Bayes theorem is to understand the theory's definition of probability. *Probability* in Bayesian terms means a measure of the belief an observer has about an outcome (Doya & Ishii, 2007). This belief is updated when the outcome arrives (Doya & Ishii, 2007). Probabilities can vary from zero to one. Loosely speaking, if the observer did not expect a certain event to occur, its probability was zero or near zero. This is because here, probability reflects a belief and not necessarily the true frequency of occurrence. To give an example, most European observers in the 18th century would have ascribed a very low probability to the event of a mammal laying eggs. Nevertheless, platypodes have been laying eggs in Australia long before the European observers arrived. Hence, a frequent event can have a very low probability in the above-defined sense. (But note that it can be shown, that if the events are observed, the probability eventually starts to mirror the true state in the environment (Friston, 2002; 2005)). Events that violate expectations, being observed even though they had been ascribed a low probability, are *surprising* – and cause an updating of beliefs. Many events co-occur reliably, thus each observation has a probability of its own, but if observations are not independent, because one event renders another more or less likely, then this is captured in their conditional probability. Take for example the probability of an animal laying eggs and the probability of an animal being a bird. The conditional probability of the animal being a bird, given it lays eggs is considerable[2]. The probability of an animal laying eggs, given it is a bird is high. Thus, in one believer, these probabilities are not independent. The immigrants to Australia would have to reverse their probability of an animal being a bird if it lays eggs. However, the probability of an animal laying eggs, given it's a bird could remain stable.

$$P(hypothesis|data) = \frac{P(data|hypothesis)\, x\, P(hypothesis)}{P(data)}$$ [3]

$$Posterior\, probability = \frac{generative\, model\, x\, prior}{normalizing\, denominator}$$

The Bayes theorem can also be applied to hypothesis testing. Here we assume the hypothesis that the probability of an animal laying eggs is low, given it's a mammal. Imagine the observer wanted to infer whether the duck-billed animal they saw was a mammal. Under the hypothesis that the animal is a mammal, the belief that it was to be observed laying eggs would be very low. The probability that the animal will lay eggs given it is a mammal would also be low; this description of how likely it is to make a certain observation given the hypothesis is called the *generative model*. The assumed prior probability of the animal being a mammal could be high and the probability of laying eggs could in itself also be high, given the occurrence of egg-laying animals that is taken into account. Lastly, the aspect of laying eggs is taken into account, captured in the normalizing denominator. Psychologists often refer the normalizing denominator as

[2] For biological precision: the probability for an animal being a bird given it lays eggs would only be *high* if only vertebrates are considered. A very large number invertebrate "animals" lay eggs that are not birds, eg., insects, nematodes, etc...,; to keep matters simple, I will use the term *animal* while *vertebrates* would be more precise,but take into account that the probability for an animal being a bird given it lays eggs would is in fact not *high*.

[3] This formula is slightly different from what psychologists or statisticians use to determine the likelihood of an experimental hypothesis (H1) tested against a null hypothesis H0). The formula used to decide what the odds of a valid H1 would be $p(H_1 | D) = \frac{p(H_1)p(D | H_1)}{p(H_1)p(D | H_1) + p(H_0)p(D | H_0)}$ (Gigerenzer & Hoffrage, 1995).

the baserate (Gigerenzer & Hoffrage, 1995). This denominator helps not to overestimate the *posterior probability* (Gigerenzer & Hoffrage, 1995), which captures the probability of the hypothesis given the observation and the estimation whereof is the goal of inference. Very loosely speaking, if an event has a high probability in itself, its observation is likely to depend on the height of its baserate and not on the presence of another observed event.

$$P(mammal|eggs) = \frac{P(eggs|mammal)\, x\, P(mammal)}{P(eggs)}$$

Observation of the fact that the animal does lay eggs could lead to estimating the probability of other hypotheses (“It's a bird.”, “It’s a fish.”, “It’s a snail.”, etc.). Here, we would expect that the likelihood of the different hypothesis shows variation; the visual evidence would probably lead to a priority in applying the generative models that concern vertebrates. The process would also lead to a reversion of the posterior probability (P(mammalleggs)). Reversing the posterior probability means that the posterior probability that an animal is a mammal given it lays eggs would be slightly enhanced. This process, which changes the generative model (or *likelihood*) and leads thus to a different estimation of the posterior probability on the next observation, describes Bayesian inference.

### *1.3.1 The Influence of Information Theory*

The *information* that we derive from one observation can be described as the *surprise* (Friston, 2010; Friston, Mattout, & Kilner, 2011; Strange, Duggins, Penny, Dolan, & Friston, 2005; Luce, 2003; Shannon & Weaver, 1949). Bayesian inference encompasses the principle of updating beliefs given surprises (Friston, 2002). Surprise is a term originally derived from information theory (Shannon & Weaver, 1949). As I will discuss later, there may be different ways surprise is dealt with by the observer. Giving an anecdotal example: when the platypus was first described in European journals, many believed that they were impositions (see Hall, 1999, for a review). Moreover, as soon as it had been established that platypus produce milk to nourish their offspring, reports that they lay eggs were simply discarded by the scientific community (Hall, 1999).

### *1.3.2 Hypotheses on the 'True State of the World'*

An influential proposition about perception, which dates back to Helmholtz (Helmholtz, 1866), proposes that the brain does not have direct access to the true state of the world. It experiences internal states, i.e. activity patterns, which accord more or less to external states. Thus the brain has to use a mechanism to infer from these (sensory) activity patterns what the state of the external world is and thus create its perceptions. We can say that perception pertains to testing hypotheses on the causes of neural activation. Current theories propose that the perception is an inferential Bayesian mechanism that attributes an external state to a specific neural activity pattern. I will call the external states *causes*[4]. A chair in the visual field of the observer, in combination with the light and shadow in the room, the luminance of the material of the chair, or the speed the chair moves by, if it moves at all, are examples of attributes that make up the cause. This cause leads to sensory activity patterns in the brain, which are elsewhere discussed as sensations, or sensory data (Friston, 2005). A certain pattern of neuronal activity in the visual cortex may correspond to a curb on the chair. It may appear as if the brain would be able to invert the relation between causes and activity

[4] Please be aware that 'cause' could likewise be e.g. an earthquake or a light touch on the shoulder and is not limited to distal events.

patterns and deduce what they were caused by. However, apart from other reasons that I will discuss in the chapter *Predictive Coding in Action Observation (1.4.1)*, inverting is not necessarily an optimal, maybe not even sufficient mechanism for perception (Kersten, Mamassian, & Yuille, 2004; Knill & Pouget, 2004). The information the brain can derive from our environment is often ambiguous and the brain meets with less than optimal circumstances. An object may be partly hidden by another. The activity pattern the partly hidden object causes would then not exactly match the stored 'activity pattern corresponding to chairs (Kersten, et al., 2004). The brain has to deal with these kinds of uncertainties to perceive the environment. The idea is that the brain behaves like an inferential Bayesian 'machine' (Friston, 2002; Knill & Pouget, 2004). It generates models on expected causes and according activity patterns. It has stored *generative models* that encompass the probability of the sensory activity given the cause. Perception pertains to calculating the probability of the cause, given the activity pattern. To that end, the activations that result from the external causes (*unpredicted input activity*) are compared against the predictions of the generative model (*model-induced activity* patterns).

The predictive coding account is the framework that has pursued the idea of the brain as a Bayesian inference machine to the largest extent and I will therefore explain the underlying concept as it is discussed in the predictive coding account. Note that I will use the terms cause, model-induced activity pattern, unpredicted input activity and generative model as outlaid above. I thereby digress from the predictive coding literature. Usually, the predictive coding literature uses the term 'sensations', or 'sensory data' to refer to unpredicted input activity patterns; sometimes the terms 'cause' and 'generative model' are used in a confusing manner, too (Friston, 2005).

### *1.3.3 Bayesian Model Adaptation: Surprise and Shannon Entropy*

Bayesian learning is set up to result, as previously mentioned, in a number of models and their ascribed probabilities very close to the distribution of causes in the real world (Friston, 2010, 2002). The probabilities that are inferred in Bayesian terms, i.e., using the probability formula above, can be used to derive at information theoretical values (Shannon & Weaver, 1949). According to information theory, each observation can be defined in the quantity of information it contains (Baldi & Itti, 2010; Doya & Ishii, 2007; Shannon & Weaver, 1949). If an event is fully predicted, it contains very little information; it is not surprising. On the contrary, events that were not predicted contain a lot of information and are very surprising. To resurrect the above example, the fact that a mammal may lay eggs is surprising and very informative, as it changes our concept of mammals. Mathematically surprise ($I(x_i)$) is described as:

$$I(x_i) = -\ln p(x_i)$$

(Baldi & Itti, 2010; Doya & Ishii, 2007; Strange, Duggins, Penny, Dolan, & Friston, 2005; Shannon & Weaver, 1949) and known in statistics as the negative log (-ln) evidence ($p(x_i)$) (Friston, 2010). Another important construct that describes the characteristics of the information derived from perception is *Shannon entropy*. Shannon entropy is again a term derived from information theory (Shannon & Weaver, 1949, but see Luce, 2003) and describes the average surprise in a series of observations (Doya & Ishii, 2007). Shannon entropy (H) is therefore mathematically described as:

$$H(x_i) = \sum_{i-k} -p(x_i)\, x\, \ln p(x_i)$$

(Doya & Ishii, 2007; Strange, et al., 2005; Luce, 2003). Thus, entropy (H) is calculated as the negative probability one outcome ($-p(x_i)$) multiplied with the logarithm of the probability of the same outcome (log p($x_i$)) summed (∑) over all possible outcomes ($_{i\text{-}k}$: that which occurred and all those known outcomes that didn't occur).

I will hereafter use the term *entropy* to denominate Shannon entropy; entropy as defined in the laws of thermodynamics is never referred to in the current work. If all observations are equally likely in that they appear equally often, each event is surprising, as it cannot be predicted (Doya & Ishii, 2007). This is the setup of the highest entropy. If entropy is large, each event is informative (Friston, 2010; Doya & Ishii, 2007; Shannon & Weaver, 1948). If one event is common and one event uncommon, only the uncommon event elicits surprise. Since this event is rare, there is little overall surprise and thus the entropy in this setup is comparatively lower than that of the previous example (Friston, 2010; Doya & Ishii, 2007; Shannon & Weaver, 1949). Frequent surprises determine large entropy, but rare surprises are in themselves more surprising. Importantly, in psychology the concept of entropy has been understood as a mathematical description of the uncertainty the observer experiences. A lot of unpredictability, as captured in high entropy, equals high subjective uncertainty (Luce, 2003; Laming, 2001). Learning can importantly be described as a reduction in uncertainty (Laming, 2001).

Different concepts of surprise, that describe in how far surprise changes the predictions of a model have been implemented. Surprising observations occur, leading to high uncertainty (hence entropy) that becomes lower if the internal beliefs are revised for example as suggested by the Kullback-Leibler Divergence (KLD). (As KLD is not central to the following experiments, I will refrain from going into detail, but attach a short description in the Appendix). For the remainder of the discussion, it suffices to know that the KLD can be used to calculate the maximum likelihood of an internal model using a least-mean square estimate as known from statistics (Doya & Ishii 2007).

It is supposed that this calculation at least roughly corresponds to the updating mechanism that underlies learning from surprising perceptions. The model of course also contains a variable that reflects potential noise (Doya & Ishii, 2007) that could relate to eg., ambiguities in the environment or noisy neural transmission.

Let us for the last time return to the rather informal platypus example. For the Europeans in Europe, the prior probabilities concerning mammals' attributes in the old world must have been rather close to the real distribution of characteristics. The reports of platypus thus met a state of low entropy. While the report itself was surprising, the low entropy itself may have lead to no or only the slightest change in posterior probability. According to Friston, states of low entropy do not impose the pressure to update beliefs (Friston, 2010). For the pioneers arriving in Australia, however, even though they must have had similar prior probabilities on the concept of mammals, meeting with the platypus, spiny anteaters, and marsupials, e.g. kangaroos and koala bears, must have contained so many surprises, that they most likely experienced high uncertainty, and revised their concepts rapidly.

## 1.4 Predictive Coding

As previously mentioned, predictive coding draws on the estimation of probabilities from the Bayes theorem and the information theoretic constructs to explain how the brain derives at perceptions. The key assumption of the predictive coding account is that the brain is organized hierarchically (Friston, 2005; Kiebel, Daunizeau, & Friston, 2008). According to Friston (2005) hierarchy discribes that "supraordinate causes induce and moderate changes in subordinate causes" (Friston, 2005, p 822, ll. 28-29). This difficult explanation is easier to understand when regarding the neuroanatomically based account by Mesulam (Mesulam, 1998), which Friston (2005) used as a basis of his hierarchical description of predictive coding. The visual system, for example, is understood to be organized in part as a linear hierarchical structure with a stepwise gradation from simple to complex representations of its input (Mesulam, 1998) However, this account demands further specification. While V1 projects to V2, V1 and V2 give rise to parallel projections to numerous peristriate association areas (Mesulam, 1998). The Mesulam account (1998) describes the hierarchy as the existence of projections from primary sensory, to upstream unimodal areas, that later reach downstream unimodal, then heteromodal and lastly paralimbic and limbic areas. But this stepwise gradation is only the main projection pathway but not exclusive: some projections cross levels (Mesulam, 1998). Thus, while the hierarchy contains projections that show a gradation to more and more integrative structures, it also contains parallel projections. The essence that remains of the hierarchical account is the existence of backward and forward projections, not precluding eg., lateral projections (Friston, 2005).

Excluding lateral projections for the sake of simplicity, predictive coding describes, that at each level of this cortical hierarchy, a generative model predicts the activity at the level below that corresponds to the assumed cause (Friston, 2002, 2005, 2010; Huang & Rao, 2011; Rao & Ballard, 1999). These model predictions are sent ‘backwards’ to the next lower level, where they result in model-induced activity that corresponds to a representation of the probability of the modelled cause at this level (Friston 2005; Huang & Rao, 2011; Kersten et al., 2004). Friston (2005) proposed that different neural populations code for the unpredicted input activity from the level below and the model-induced activation that results from back-projections from the higher level. The model-induced activation derived from back-projections is compared to the unpredicted input activity at the respective level, derived from forward-projections of the level below; the mismatch of model-induced activation and unpredicted input from the level below is transferred via forward connections to the next higher level (Friston, 2005). Such a mismatch is called the *prediction error* (Friston, 2005, 2002). The prediction error can lead to adjustment of the generative model at the higher level of the cortical hierarchy (Friston, 2005). This coding pertains to the concept of sparse coding or redundancy reduction (Huang & Rao, 2011), since all predicted inputs are filtered and elicit no additional activation. In sum, we find that the calculation at each level of the cortical hierarchy pertains to calculating the probability of the cause assumed by the model, given the data. This description equals the left side in the above-formulated Bayes theorem; it means calculating the posterior probability of the model. So the calculation the brain has to make to achieve perception is equal to the right side of the equation and combines the likelihood of an activity pattern given the model, the prior probability of the model and the base rate of the activity pattern (Doya & Ishii, 2007).

In a seminal article, Rao and Ballard (1999) used a computer implementation of the predictive coding account to show how the phenomenon of end-stopping could occur. End-stopping describes the characteristic of cells in the visual cortex that fire to a stimulus consisting of a line with a certain orientation, but fire less if a longer line of the same orientation is presented (Finlay, Schiller, & Volman, 1976; Hubel & Wiesel, 1965; Schiller, Finlay, & Volman, 1976). The Rao and Ballard computer model was trained with naturalistic images; thus, their model had come to expect lines of a length that extended beyond the receptive fields of neurons in V1. The expectation of a line length would therefore increase prior activity in area V2, which contains neurons with large receptive fields. The activity of neurons in V1, which displays smaller receptive fields, would therefore be fully predicted by activity in V2, silencing activity in V1. In other words, there was no prediction error that would need to be conveyed from V1 to

V2. As the model, however, predicts lines to be of a certain length, in accordance with the stored representations it has acquired during training with naturalistic images, uncommonly short lines are not predicted by V2, causing mismatch activity in V1 (Friston, 2005; Rao & Ballard, 1999).

It has been suggested that a well-adapted observer will experience less surprise than an ill-trained observer (Friston, 2002). In fact, trained with naturalistic images, the model would be a well-adapted observer in the real world, but not in an artificial setting that employs un-naturalistic short lines to test responses to line orientation. But this proposal of trained adaptation must be regarded carefully. Consider environments of high entropy. Even a well-trained observer will experience a lot of surprise. However, the well-trained observer will come to expect these surprises based on the frequent recent surprises, i.e., will expect large entropy. It may be more succinct to conclude that a well-trained observer experiences less surprise than an ill-trained observer *in a predictable environment*; and, in addition, that a well-trained observer is less surprised at each surprise *in an unpredictable environment*.

Friston and colleagues further informally proposed that an observer could minimize the surprise she experiences, if she moved to a dark room and closed her eyes, according to the authors “a nice description of going to bed” (Friston, Daunizeau, & Kiebel, 2009). However, when we open our eyes in the morning and switch on the light, we are usually not massively overwhelmed by surprise. Why is that? I propose that prediction does not only precede perception, but can also precede what has been called sensation, activity patterns corresponding to the visual input before subjected to internal representations. There is a temporal autocorrelation of perceptions. Thus, at nearly each moment in time, we have a fair expectation of the sensory (and motor) activity pattern that will arrive if we move our heads, or eyes, or if we get up and walk into the next room. Of course, the reliability and amount of these predictive models depend on our previous experience with a certain environment – we have a better model of what to expect in our own bedroom than what to expect in the zoo (potentially not only a less-well known environment, but also one displaying higher entropy than the bedroom). However, in the described cases, predictions exist prior to the visual input concerned, based on the environment the observer is in. Prior expectations are as I have described central to Bayesian inference and has been proposed to be revealed in the spontaneous activity of neural populations (Fiser, Berkes, Orbán, & Lengyel, 2010).

### *1.4.1 Predictive Coding in Action Observation*

Predictive coding has been applied to explain the neuronal activity during action observation as measured for example using fMRI. Action observation and imagination lead to activity in the cortical motor system. The main components of the cortical motor system are the premotor and the posterior parietal cortex (Jeannerod, 1995.), but also temporal or even occipito-temporo-parietal areas (Beauchamp & Martin, 2007; Kilner, Friston, & Frith, 2007) The predictive coding account of action observation draws on the idea of forward models in action and motor control, as I have described in the chapter *Early Forward Models*. However, the distinction between forward and inverse models is aborted in predictive coding (Friston, 2011). The first reason is a critique of the hidden states that inverse models demand. But more importantly, an account that relies on inversion demands revertible models. However, an action that is performed in one context, in the Jekyll and Hyde example manipulation of a scalpel to cure in an operating theater (Jacob & Jeannerod, 2005; Kilner, et al., 2007), could aim at a different action goal than manipulation of a scalpel in the street in the second the example (the goal could be to hurt someone; Jacob & Jeannerod, 2005; Kilner, et al., 2007). Inversion of the underlying forward model of the manipulation would not necessarily allow discrimination of the action goal. The top-down, or backprojections in predictive coding are not efference copies of the motor command, but a prediction of the activity pattern in sensory cortices that concurs with the motor command (Friston, 2011). The timecourse or order of activity increase in all areas from the primary visual cortices to premotor or even prefrontal sites (Csibra, 2007; Jacob & Jeannerod, 2005) that become activated during action observation not entirely known. However, it is assumed that ultimately, a generative model at the highest level of abstraction predicts the neural activity at the level below and so forth. Parenthetically, it seems rather unclear what the meaning of ”highest area in the brain” in predictive coding accounts could be.

In any case, the predictive coding account proposes that as soon as a high level representation of the action exists, its predictions concerning activity at the next lower level are back-projected. This, of course, accords to all levels of the hierarchy that derive through the described mechanism of mismatch detection and model adjustment at the most likely perception. An activity pattern corresponding to the visual input of biological motion, for example, as long as it’s not predicted by the current modell causes a prediction error in the next higher level of the hierarchy, where it activates a representation that predicts what the activity reflecting the biological motion should be like, if it can be explained by this representation. This process of model adjustment until no mismatch occurs explains hence the activation of the cortical motor network in action observation.

### *1.4.2 The Predictive Coding Explanation for Evoked Brain Activity*

Predictive coding has the benefit of being applicable to different levels of neural coding, ranging from primary sensory to unimodal or integrative areas and beyond. It can explain how a visual input is inferred to be caused by a hand in our visual field, but, using an example discussed by Jacob and Jeannerod (2005) as well as Kilner, Friston, and Frith (2007), it can also be used on a higher level to infer whether the trajectories of the hand relate either to curing someone or hurting someone with a scalpel. Predictive coding has found wide adaptation in describing cortical activity, in fact, an influential theoretical paper on the matter is entitled “A theory of cortical responses” (Friston, 2005). However, very few attempts have been made to relate predictive coding to subcortical responses (but cf. Friston, et al., 2009; Huang & Rao, 2011). This is somewhat surprising, given that the concept of prediction errors in the context of reward has been researched extensively in the midbrain dopaminergic nuclei and the striatum. This extensive research is based on the temporal-difference algorithm, which I will come to describe shortly. One reason for the lack of investigation of subcortical prediction errors in perception may be, that while the temporal-difference algorithm (*TD-algorithm*; Montague, Dayan, & Sejnowski, 1996) is explicitly meant to learn to predict future states, predictive coding is not regarded to yield prediction on future states but only be concerned with the predictions one level of cortical hierarchy makes for activity on the next lower level (Kilner, et al., 2007). The last argument however, can be disputed. One critique is that if predictive coding is used to perceive events, a temporal dimension is necessary. Secondly, if predictive coding in action observation derives inferences on intentions, for example drinking a glass of water. The predictions of intentions to the next lower level should consist also in future action steps, e.g. taking the glass, turning in the tap, filling the glass, etc., because these steps should be encompassed in the representation of filling a glass of water. In other words, if one of these steps did not occur, this could mean that the generative model at the "highest level" needs adjustment. In sum, the existence of a generative model of intentions, as proposed by Kilner and colleagues in the Jekyll & Hyde example (2007) determines a predictive process encompassing a temporal component. Parenthetically, a serial or temporal account of predictive coding can be found in the works of Mehta (2001).

## 1.5 Reward Related Prediction Errors

The TD-algorithm of reward related learning explains how prediction error based learning enables the brain to predict the occurrence of reward and associate certain behaviours with reward, or the omission of reward (Schultz, 2000; Schultz & Dickinson, 2000; Schultz, Dayan, & Montague, 1997). The TD-algorithm owes part of its acclaim to the fact that has been used to successfully model the response of midbrain-dopaminergic neurons (Schultz et al., 1997). The underlying idea of learning from prediction errors was also present in predecessors of the TD-algorithm, as will be outlined in the following.

### *1.5.1 Surprise Incites Learning – the Rescorla-Wagner Model*

Not the first (cf. Kamin, 1969) but one of the most influential predecessors of TD that explained how unpredicted events incite learning was the Rescorla-Wagner learning rule (Rescorla & Wagner, 1972). Based on their experiments on Pavlovian fear conditioning, the authors postulated that an organism would only learn as long as events violate its expectations (Rescorla & Wagner, 1972).

A number of facts make the Rescorla-Wagner rule noteworthy, regardless the large number of mathematically more refined later models. First of all, the use of the term surprise, which we incidentally already find in the work of Kamin (1969), appeals due to its clear psychological meaning. It is also the term that has re-emerged in very recent models of brain functioning such as the predictive coding account (Baldi & Itti, 2010; Friston, 2010; Itti & Baldi, 2005). The model also features incremental learning. The more often a conditional stimulus - unconditional stimulus pairing has been witnessed, the closer is the predictive capacity of the conditional stimulus to the real rate of occurrence of the unconditional stimulus.

## 1.6 Prediction Error Driven Learning: the TD-algorithm

The TD-algorithm of reward related learning is very similar to the Rescorla-Wagner learning rule, but is has the benefit of making temporal predictions. That means, that the model does not only learn to predict the occurrence of sensory states (reflecting the sensory consequence for example of perceiving events, or of conducting an action), but it also learns when these states will occur (Montague, et al., 1996). The term sensory state is not directly related to the term sensation, or sensory data in predictive coding (Friston, 2005), but concerns activity patterns spread over cortical or subcortical components, reflecting all current input, regardless predictions. The reception of reward is also a sensory state for TD, albeit of a somewhat different quality (Montague, et al., 1996). For the matter of temporal prediction, time is represented in discrete time steps. The easiest description of a TD learning algorithm is that it learns for each state how much reward is to be expected in a (usually undefined) number of future states (Montague, et al., 1996). For each transition to the next time step into the future, the model compares the reward it received at that time step plus the reward the current state predicts, with the predictions of reward for all future time steps that was current at the last time step (Montague, et al., 1996).

Two facts are of major importance for the TD-algorithm. The first is, that it will eventually come to fully predict the reward, hence, the prediction error will cease (Schultz, et al., 1997). The second aspect is that the prediction error will slowly transfer to the first sensory state that reliably precedes the reward in a temporally fixed manner, i.e. a predictive sensory cue (Schultz, et al., 1997). The prediction error will then stop to propagate backwards. I will include a more detailed account of the TD-algorithm in the Appendix, but the last mentioned aspects will suffice to understand the following discussion. The acclaim of the TD-algorithm is largely due to the fact that it has been successfully applied to the response of midbrain dopaminergic cells (Schultz et al., 1997). The cells in the primate ventral tegmental area (VTA) have been shown to fulfil the suppositions the TD-algorithm makes for a neuronal population coding for a prediction error. Among these fulfilled expectations were that the prediction error will eventually occur to the predictive cue, that it will cease to occur for the predicted reward, and that an omitted predicted reward decreases cell firing (Schultz, 2000; Schultz et al., 1997; Schultz, Apicella, & Ljungberg, 1993; Suri, 2002).

The results from single-cell recordings in non-human primates have later been transferred to research in the human brain, and fMRI has been used to investigate the response of the human dopaminergic system (O'Doherty, Buchanan, Seymour, & Dolan, 2006; O'Doherty et al., 2004). However valuable invasive research methods in animals may be, interpretation of the results should not neglect an important fact: animals will cooperate for reward and to avoid punishment. Testing animals in a reward-free environment devoid of incentive punishment is nearly impossible. Findings that relate to reward in the animal may actually not be reward-dependent when investigated in the human. In line with this proposal, a number of authors have proposed that the midbrain dopaminergic system is in fact not responsive to reward, but to all *salient* events (Horvitz, 2000; Redgrave & Gurney, 2006). Given that the TD-algorithm is in fact an algorithm that learns to predict states (related to sensory or motor input and output), it is possible that brain areas that have shown reward-related prediction errors in animal research are in fact responsive to prediction errors per se. I have mentioned that research on midbrain dopaminergic nuclei played a pivotal role in the success of the TD-algorithm. Moreover, dopamine is understood to fulfil a number of learning (Reynolds & Wickens, 2002) and decision-making (Frank & Claus, 2002) related function in the basal ganglia. I will therefore give a short summary of important research results on this neurotransmitter.

## 1.7 Dopamine

Dopamine is a monoamine neurotransmitter. The substantia nigra and ventral tegmental area (VTA) are the primary source of dopamine in the brain. The pathways have recently been reviewed by the group of Bjoerklund (Bjoerklund & Dunnett, 2007), one of the pioneers of research on dopaminergic projections in the 1970s and 1980s (Bjoerklund & Dunnett, 2007; Lindvall, Bjoerklund, & Divac, 1978; Lindvall, Bjoerklund, & Skagerberg, 1984). Three major pathways project to the forebrain, the mesocortical, the mesolimbic and the mesostriatal pathway. The latter is better known as nigrostriatal pathway, but the nomenclature seems to undergo some change as information on the origin of this pathway increases (see Bjoerklund, 2007 for review). The mesocortical pathway projects mainly to the prefrontal and to a degree to the premotor cortex (Bjoerklund, 2007; Gaspar, Stepniewska, & Kaas, 1992; Le Moal & Simon, 1991). The mesolimbic pathway targets the amygdala, olfactory tubercle, the nucleus accumbens and septum (Bjoerklund, 2007; Le Moal & Simon, 1991) and the mesostriatal pathway sends dense projections to the dorsal striatum, i.e., caudate nucleus, putamen and globus pallidus (Bédard, Larochelle, Parent, & Poirier, 1969; Haber, 2003).

Dopamine binds to five receptor subtypes, D1 to D5 (Missale, Nash, Robinson, Jaber, & Caron, 1998). Receptor subtypes D1 and D5 are usually subsumed as receptors of the D1-like family type (Missale et al., 1998) and I will refer to these receptors simply as D1 receptors. D2, D3 and D4 receptors are usually subsumed as the D2-like family type, (Missale et al., 1998) and I will refer to them as D2 receptors.

A role for dopamine has been proposed for a large number of functions, e.g. movement (Lindvall et al., 1990), working memory (Durstewitz, Seamans, & Sejnowski, 2000), attention (Rose, Schiffer, Dittrich, & Gunturkun, 2010), reward-related learning (Schultz et al., 1997), 'feelings' of hedonia (Gardner & Lowinson, 1993; but see Berridge & Robinson, 1998), and novelty responses (Horvitz, 2000; Redgrave & Gurney, 2006)

Dopaminergic malfunction is involved for example in Parkinson's Disease (PD), attention deficit hyperactivity disorder, Schizophrenia, and drug addiction (e.g. Barbeau, 1970; Berke & Hyman, 2000; Bernheimer, Birkmayer, Hornykiewicz, Jellinger, & Seitelberger, 1973; Chouinard & Jones, 1978; Dagher & Robbins, 2009; Gardner & Lowinson, 1993; Kelley, 2004; Levy & Swanson, 2001; Lindvall et al., 1990; Schultz, 2007, for reviews)

## 1.8 The Basal Ganglia

The basal ganglia derive their name from *basal* - bottom or deep, and *ganglia* - collection of nerve cell, which adheres to their location in the midbrain. The basal ganglia encompass the following nuclei: caudate nucleus, putamen, nucleus accumbens (N. Acc), globus pallidus externa (GPe), globus pallidus interna (GPi), subthalamic nucleus (STN), substantia nigra pars reticulata (SNr), substantia nigra pars compacta (SNc). An important functionally related structure is the ventral tegmental area (VTA).

The striatum (*striatus* = grooved) the largest nucleus of the basal ganglia. The primate striatum can be subdivided in three separate nuclei, the putamen (putamen = shell), the caudate nucleus (cauda = tail) and the nucleus accumbens (accumbere = to lie/lean adjacent to). Putamen and caudate nucleus are separated by the internal capsule. The internal capsule in fact gives the striatum is distinctive grooved look that inspired the name corpus striatum, given by Thomas Willis (cf. Meyer & Hierons, 1964). The striatum itself can be divided into the dorsal striatum (caudate nucleus and putamen) and the ventral striatum (nucleus accumbens). (For more detailed descriptions of basal ganglia anatomy, see Bolam, Brown, Moss, & Magill, 2009; Meyer & Hierons, 1964; Parent & Hazrati, 1995a/b; Saint-Cyr, 2003; Smith, Bevan, Shink, & Bolam, 1998.) A structure that has been related increasingly to basal ganglia function is the habenula (habena = reigns) in the epithalamus (Hikosaka, Sesack, Lecourtier, & Shepard, 2008; Lecourtier & Kelly, 2007; Matsumoto & Hikosaka, 2007).

I will discuss two anatomical aspects of the striatum that are of importance for the hypotheses that guided the experiments of my thesis, namely its interconnectedness and its dopaminergic innervation. I will then progress to its putative functions.

The striatum as the input structure to the basal ganglia shows a remarkable pattern of connection that has been matter of research and heated debate for more than 25 years (Alexander, DeLong, & Strick, 1986; Haber, 2003; Parent & Hazrati, 1995a; Selemon & Goldman-Rakic, 1985). In describing the anatomical connections of the basal ganglia (and thus the striatum) it is important to distinguish between two important concepts.

The first is that of cortico-basal ganglia-thalamo-cortical loops (Alexander, et al., 1986; Haber, 2003; Parent & Hazrati, 1995a; Selemon & Goldman-Rakic, 1985). The second is that of the basal ganglia pathways (Albin, Young, Penney, Roger, & Young, 1989; Gerfen & Surmeier, 2011; Haber, 2003; Smith, et al., 1998). I will shortly give an overview of the loop concept and later explain the pathway concept.

### *1.8.1 Cortico-Basal Ganglia-Thalamo-Cortical Loops*

The characteristic of the cortico-basal ganglia-thalamo-cortical loops as described by Alexander and colleagues (1985) is that of partially open loops. The principle can be explained thus: certain cortical areas project to the same area of the striatum. These projections to the striatum give rise to even more converged projection zones in the output nuclei of the striatum, the GPi and SNr. The information is then transferred via the thalamus to one of the cortical input regions. The description of partially open loops is due to the fact that while the projections of a number of regions converge on one striatal area, the backprojections from thalamus to cortex reach only one (and always the same) input region, but not all input regions. Thus, we find a closed loop for one input region (the one that is also the output region), but due to the other input regions that receive no thalamic back-projections, we call this concept partially open loops. I will describe one exemplary loop of the five originally proposed loops by Alexander and colleagues (1985) in exactly the way the authors did at the time to clarify the matter.

The motor loop has inputs from the supplementary motor area, the arcuate premotor area (that can be regarded as the monkey lateral premotor cortex), the motor cortex and the somatosensory cortex. These projections converge in the same area of the putamen. The putamen then projects to the ventrolateral GPi and caudolateral SNr. The projection from these output nuclei then reaches the thalamic nuclei ventralis lateralis pars oralis and ventralis lateralis pars medialis. The thalamic cortical projection in the motor loop reaches *only* the SMA. The same principle can be found for all ‘Alexander loops’. Of specific interest in the dorsolateral prefrontal loop, projections from the dorsolateral prefrontal cortex (dlPFC), from the posterior parietal cortex and arcuate premotor area reach the same area of the dorsolateral head of the caudate nucleus (Alexander et al., 1985), hence projection sites within one loop do not necessarily have adjacent input areas (Selemon & Goldman-Rakic, 1985).

The debate that I have earlier referred to, concerning closed, open and partially open loops, sparked at the idea of convergence of information from different brain areas in the striatum (Haber, 2003; Parent & Hazrati, 1995a; Selemon & Golman-Rakic, 1985). It is still not clear whether areas from different loops also connect in the striatum, but the interdigitation of projections could enable dendritic arborization to lead to an information transfer between loops (Haber, 2003; Parent & Hazrati, 1995a; Selemon & Golman-Rakic, 1985). It has been suggested that cortical input projects to different kinds of compartments in the striatum (Parent & Hazrati, 1995a). One type of compartment in the striatum are the striosomes or patches, the other the extrastriosomal matrix, that surrounds the patches (Smith et al., 1998). The extrastriosomal matrix contains output nuclei, the matrisomes (Parent & Hazrati, 1995a). This distinction is relevant concerning the shaping of associations between different cortical input areas that I will discuss next. It was proposed that the matrisomes act as templates wherein associations between the activation pattern of different cortical areas can be ‘chunked’ together, i.e., associated with each other (Graybiel, 1998).

Another possible mechanism for information transfer between the projection sites of different loops are subcortical loops through the basal ganglia (Haber, 2003; McHaffie, Stanford, Stein, Coizet, & Redgrave, 2005), especially striato-nigral-striatal loops

(Haber, 2003). The last point deserves clarification as I have mentioned dopamine in the context of reward-related learning and prediction errors and will come to talk about it more extensively in connection with the basal ganglia pathways.

In very easy terms, the ventromedial striatum receives input from a small dopaminergic midbrain region but sends projections to a large midbrain region. In contrast, the dorsolateral striatum receives input from a large midbrain dopaminergic region, but projects only to the ventral regions of the midbrain (Haber, 2003; cf. Bjoerklund, 2007). The ventromedial striatum could thus influence the dopaminergic input to the dorsolateral striatum. Since different cortico-basal ganglia-thalamo-cortical loops seem to traverse the ventromedial and dorsolateral striatum (Alexander et al., 1985), this mechanism could offer a way for different cortico-basal ganglia-thalamo-cortical loops to influence each other via mediation of the striato-nigro-striatal loops. In fact, the mechanism could be of tremendous importance to learning: dopaminergic projections from the substantia nigra to the striatum change the synaptic plasticity in the striatum. Dopamine binding at D1 receptors furthers long-term potentiation (LTP), while dopamine binding to D2 receptors inhibits long term potentiation (Reynolds & Wickens, 2002; Wickens, Horvitz, Costa, & Killcross, 2007). Thus, activity in the ventromedial striatum could influence the dorsolateral striatum via projections to the substantia nigra, resulting in the learning of new associations. In contrast to the dorsolateral striatum, the ventromedial striatum is particularly associated with reward-based learning (O’Doherty et al., 2004). The striato-nigro-striatal loop could thus provide a mechanism that could potentially allow medial frontal and oribitofrontal areas to influence learning for example of motor responses, by fostering learning of associations in the motor cortico-basal ganglia-thalamo-cortical loop.

### *1.8.2 Basal Ganglia Pathways*

The original notion was that of two basal ganglia pathways, one to enable a reaction, and one to suppress reactions (Albin et al., 1989; Bischoff-Grethe, Crowley, & Arbib, 2002; Frank & Claus, 2006). Current theories assume a more complex organization, with internal loops within the pathways and an additional so-called hyperdirect pathway (Frank, Samanta, Moustafa, & Sherman, 2007). I will describe the original findings (Albin et al., 1989) and give a short summary of the proposed changes.

The two basal ganglia pathways are called the direct and indirect pathway. Each starts in the striatum in cells that either express the D1 or D2 receptors. The direct pathway consists of projections from the medium spiny neurons that are characterized by expressing D1 receptors (Bolam, et al., 2009). This projection reaches the output structures, that is the GPi and SNr, directly. This projection is GABAergic and thus inhibits the GPi and SNr. The GPi and SNr send projections to the thalamus that are likewise GABAergic. The inhibition of the GPi and SNr via the D1 receptor-expressing striatal neurons thus dampens the inhibitive projections from GPi and SNr to the thalamus, disinhibiting the thalamus. The indirect pathway has the opposite function. D2 receptor-expressing medium spiny projection neurons in the striatum send GABAergic projections to the GPe. The GPe has GABAergic projections to the STN. The STN in turn sends excitatory projections to the GPi and SNr. If the D2 receptor-expressing striatal cells are activated they inhibit the GPe. The inhibition of the GPe leads to a disinhibition of the STN. If the STN is thus activated, its excitatory projections to the GPi and SNr lead to heightened activity in these output nuclei. The output nuclei's activity inhibits the thalamus via GABAergic projections. In addition, the STN also sends backprojections to the GPe. The important fact is, that activation of the D2 receptors leads to inhibition of the indirect pathway. Dopamine activates the direct pathway via binding to D1 receptors and inhibits the indirect pathway via binding to the D2 receptors (Bolam et al., 2009). Dopamine has thus, in short, a disinhibiting function on the thalamus, and hence, the cortex.

As I have described previously, D1 receptor activation leads to long-term potentiation, while D2 receptor activation prevents long-term potentiation. If the representation of cortical activity transferred via any loop in the striatum is accompanied by a dopamine burst, this will first of all lead to a “go” response, for example the execution of a represented motor command (Smith et al., 1998). However, it will also cause the D1 receptors of the direct pathway to express long-term potentiation and the D2 receptors of the indirect pathway to show no long-term potentiation or even long-term depression (LTD). Thus, the dopamine burst will teach both pathways to make one response more likely, while concurrent alternative responses are suppressed (Frank, 2006 for a review).

There are two reasons for my giving this detailed account: Firstly, it is important not to confuse the cortico-basal ganglia-thalamo-cortical loops with the pathways. Hence, it is important to realize that the pathways are a potential part of any cortico-basal ganglia-thalamo-cortical loop (Smith et al., 1998), but do not constitute separate loops themselves. The second reason is the involvement of dopamine in the pathways. D1 and D2 receptors are associated with different loops and thus with different functions. D1 receptor binding in the striatum will lead to a disinhibition of the thalamus, increasing cortical activity. If we consider the role of the direct and indirect pathway for example in the motor loop, D1 receptor binding could thus constitute a potential “go” signal for a motor command. Activity of the D2 receptors on the other hand leads via the indirect pathway to inhibition of the thalamus and thus represents a “no-go” signal (Frank, et al., 2007). Regarding the mechanisms of LTP (and possibly LTD), D1 receptor activation also increases the synaptic strength of the representation currently held in the direct pathway.

In addition to the two pathways that I have described above, there are two major additions to the model: a direct projection from the GPe to the output nuclei and a hyperdirect pathway from the cortex via the STN to the output nuclei (Parent & Hazrati, 1995b). The discussion of the hyperdirect pathway would go beyond the limits of this thesis, but it is important to understand that the hyperdirect pathway makes it possible for the mesial prefrontal cortex to issue a global 'no-go' signal via its projections to the STN and the STNs excitation of the output nuclei. Thus, thalamic and cortical activity is inhibited and responses are prevented (Frank, et al., 2007). The mesial pre-frontal cortex can thus modulate the responses corresponding to the computations of the direct and indirect pathway.

## 1.9 Summary and Research Questions

To introduce the theoretical background of my work, I have so far described two influential models of brain function. The theory of predictive coding relates neuronal responses to states that are not predicted by any internal generative model. Predictive coding has been described as a powerful way to explain cortical responses, but apparently not been related to subcortical responses. Predictive coding, due to its Bayesian base, leads to clear hypotheses on when and how to the generative models that guide perceptions are updated. The predictive coding account claims, however, not to be related to forecasting future states. And even though it has a clear connection to the predictions of the motor system, it does not relate to a structure known to be involved in motor learning: the basal ganglia.

The temporal difference algorithm on the other hand provides an account of how we learn to predict future states. Its neural implementation has been ascribed to the dopaminergic midbrain and the striatum. The fact that the seminal studies that established neuronal responses according to the TD-algorithm were animal studies, may have led to a disregard of the possibility that prediction errors concerning future sensory events may be likewise coded for in the striatum. Both models agree on the assumption that only unpredicted events cause neuronal activity, and that the purpose of learning is to diminish the prediction error. The predictive coding account moreover borrows from information theoretic concepts to predict when learning should take place, and what type of learning can occur.

Action observation is the hallmark paradigm for eliciting emulation or prediction based on an internal model. The motor loop is the most thoroughly researched cortico-basal ganglia-thalamo-cortical loop and current theories hold that internal models of actions may be passed through this loop before they are executed (Redgrave, Prescott, & Gurney, 1999). Thus, the activity of the cortical motor system during action observation promotes the idea that internal forward models are represented in cortico-basal ganglia-thalamo-cortical loop activity. I therefore used action-observation paradigms in the following experiments to aim at three goals:

To test for perceptual, non-reward related prediction errors in the striatum.

To investigate when the brain learns to predict visual input.

To investigate the correlates of high entropy that, as such, prevents predictions of events, but allows predicting the occurrence of prediction errors.

# 2 Research Articles

## 2.1 Caudate nucleus signals for breaches of expectation in a movement observation paradigm.

Caudate article

frontiers in
HUMAN NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 08 April 2011
doi: 10.3389/fnhum.2011.00038

# Caudate nucleus signals for breaches of expectation in a movement observation paradigm

***Anne-Marike Schiffer* and Ricarda I. Schubotz***

*Motor Cognition Group, Max Planck Institute for Neurological Research, Cologne, Germany*

***Edited by:***
*Hauke R. Heekeren, Max Planck Institute for Human Development, Germany*

***Reviewed by:***
*Jeffrey M. Zacks, Washington University, USA*
*Simone Schütz-Bosbach, Max Planck Institute, Germany*

****Correspondence:***
*Anne-Marike Schiffer, Motor Cognition Group, Max Planck Institute for Neurological Research, Gleueler Straße 50, 50931 Cologne, Germany.*
*e-mail: schiffer@nf.mpg.de*

The striatum has been established as a carrier of reward-related prediction errors. This prediction error signal concerns the difference between how much reward was predicted and how much reward is gained. However, it remains to be established whether general breaches of expectation, i.e., perceptual prediction errors, are also implemented in the striatum. The current study used functional magnetic resonance imaging (fMRI) to investigate the role of caudate nucleus in breaches of expectation. Importantly, breaches were not related to the occurrence or absence of reward. Preceding the fMRI study, participants were trained to produce a sequence of whole-body movements according to auditory cues. In the fMRI session, they watched movies of a dancer producing the same sequences either according to the cue (88%) or not (12%). Caudate nucleus was activated for the prediction-violating movements. This activation was flanked by activity in posterior superior temporal sulcus, the temporo-parietal junction and adjacent angular gyrus, a network that may convey the deviating movement to caudate nucleus, while frontal areas may reflect adaptive adjustments of the current prediction. Alternative interpretations of caudate activity relating either to the saliency of breaches of expectation or to behavioral adaptation could be excluded by two control contrasts. The results foster the notion that neurons in the caudate nucleus code for a breach in expectation, and point toward a distributed network involved in detecting, signaling and adjusting behavior and expectations toward violated prediction.

**Keywords: caudate nucleus, movement observation, prediction, fMRI, internal model, biological motion, expectation, frontal lobe**

## INTRODUCTION

The striatum was once considered a site of solely motor function, but research over the last three decades has put its cognitive functions more and more into focus (e.g., Alexander et al., 1986; Saint-Cyr, 2003; Grahn et al., 2008). One prominent function of the striatum is the coding of a reward prediction error in learning. These prediction errors are triggered by reinforcement or reward in conditioning paradigms (Schultz et al., 1997, 1998; Schultz, 2000). Reward prediction errors are signified by increases in striatal firing in the presence of unexpected reward or the presence of a reward-predicting cue, or by a decrease of firing when predicted reward is omitted. The underlying notion to a reward prediction error is that the brain is capable of associating the current circumstances with a specific future state (Wolpert and Flanagan, 2001; Friston, 2010). If the future becomes present and the state is different from what was predicted, this violation of predictions causes a prediction error, which in turn incites learning.

That the brain is a "predictive machine" is a feature of many models concerned with learning, action, and perception (Rescorla and Wagner, 1972; Schultz and Dickinson, 2000 Kiebel et al., 2008; Bubic et al., 2009; Friston, 2010). In an extension of the theory of motor control (Wolpert and Flanagan, 2001), the brain's ability to constantly predict ongoing movement, be it in the motor domain or in perception, has been emphasized (Schutz-Bosbach and Prinz, 2007). Presence of prediction implies the possibility of computing prediction errors, to adjust internal models according to perceptions and thus shape the correct predictions. The exact anatomic implementation of not reward-related prediction error signals, that code for unexpected perceptions has yet to be revealed (Zacks et al., 2007). The proposed involvement of dopamine (Zacks et al., 2007) and the striatum's extensive connectivity (Alexander et al., 1986; Saint-Cyr, 2003) render it a likely candidate as a site of not only reward-related prediction errors but also more general not reward-related prediction errors.

In fact, there is some evidence that striatal firing-patterns indeed convey prediction errors that are not related to reward. The respective authors likewise used the term prediction error to describe this dorso-striatal activity (Horvitz, 2000; Schultz and Dickinson, 2000; Graybiel, 2005). For the sake of distinction, we will refer to activation that has an amplitude that varies with the amount of (expected) reward as "reward prediction errors." The prediction errors investigated in the current study, that have a positive amplitude to all violated predictions, will be called "breach of expectation" signals. This breach signal is related to a violated prediction in the simplest sense, i.e., a prediction of any given content is not fulfilled. Accordingly, increased activity toward every unexpected stimulus signifies the breach of expectation signal in dorsal striatum. Indeed, recent imaging studies in humans report caudate nucleus activity for unexpected changes in context, rules, and contingencies (Bunge et al., 2003; Delgado et al., 2005; O'Doherty et al., 2006; Badgaiyan et al., 2007; Koch et al., 2008; den Ouden et al., 2009, 2010). These activations can be broadly interpreted as corresponding to the occurrence of unpredicted stimuli. The study of Davidson et al. (2004), for example, revealed a negative response of the caudate to unexpected target omission, which could be reframed as occurrence of a unpredicted target-free stimulus.

Taken collectively, the data suggest that caudate prediction error signals are not restricted to conditioning protocols and that they do not revolve solely around the availability of reward. The empirical evidence implies the presence of breach of expectation signals in caudate nucleus when an event deviates from predictions, but there is a need to probe the assumption directly.

The lack of studies that target breach of expectation signals is surprising given not only the role they play in current computer models on the matter (Kilner et al., 2007; Kiebel et al., 2008; Friston, 2010) but the enormous relevance of correcting false assumptions to prevent possibly fatal future mistakes. The educative effect of breaches of expectation is so strong that it operates even when observing other peoples behavior. Consider the example of seeing someone being bitten by a bulldog after having tread on its paw. If you used to regard bulldogs as aggressive animals, this would not breach your expectations and not incite changes in your views on bulldogs and your behavior toward them. In other words, you wouldn't have learned anything. Now consider watching someone being bitten by a poodle after stepping on its paw. This may be a severe breach of your expectations and teach you to regard poodles more suspiciously in future and adapt your behavior toward them accordingly. This is an example of observational learning, which does not imply direct reinforcement – it also embellishes two things. The first is the importance of the severity of a breach of expectation to learning. The second is the bonus derived from valid forward models in guiding behavior.

The current study investigates whether caudate nucleus signals for breaches of expectation in a movement observation paradigm. We hypothesized that watching a dancer make a mistake in a setting of clear-cut cue-movement schedule would yield a caudate response. To keep track of the dancer's performance, participants had to register auditory cues that determined what movement the dancer was to perform next, and watch the ensuing movements. To ensure that all participants were capable of the required prediction, we subdued them to motor training, where they had to accord to the cue-movement schedule themselves.

Breaches of expectation carry two secondary attributes that could each potentially cause striatal activations. Specifically, these events are of an increased saliency and often prompt to modify ongoing behavior. Saliency can be conceived of as a function of stimulus frequency (Zink et al., 2003) and is an attribute carried of not-habituated stimuli (Redgrave and Gurney, 2006). As violated predictions were rarely encountered in the present paradigm, i.e., infrequent, they might hence elicit striatal activation due to their saliency (Horvitz, 2000; Redgrave and Gurney, 2006). Movement switches as opposed to executing one movement repeatedly have also been associated with striatal activity (Roy et al., 1993; Graybiel, 2005). Encountering violated predictions in the paradigms is related to having to switch to a new internal movement simulation to keep track of the task. Moreover, as the paradigm included extensive training, there may have been an association of movement errors with initiation of a new movement. Hence, saliency and movement switches had to be investigated separately, to ensure that these attributes of violated predictions could not account for potentially recorded striatal activity.

We employed an experimental design that allowed to test whether caudate activity actually reflects breaches of expectation (violation hypothesis – i), or is rather dependent on effects of saliency (saliency hypothesis – ii) or switching to a different behavior (change hypothesis – iii). Breaches of expectation were modeled by contrasting predicted with prediction-violating movements. In accordance with the frequency or habituation approach in the literature, we modeled saliency as a function of stimulus frequency in the immediate trial history. Initiation of a new movement was implemented in the movement observation paradigm by contrasting the cues that indicated a new upcoming movement against cues that indicated a movement repetition.

Although the present study focused on striatal responses, it was to be expected that they come along with cortical activations, as a prominent characteristic of the neostriatum is its pronounced connectivity with a large number of cortical regions and thalamic nuclei (Alexander et al., 1986; Saint-Cyr, 2003). More specifically, activation could be expected in regions related to the processing of biological motion (due to the mismatch between perceived and expected stimulus; Keysers and Perrett, 2004) and those related to attentional modulation more generally (due to the explicit instruction rendering breaches of expectation task-relevant; Corbetta and Shulman, 2002).

## MATERIALS AND METHODS

### PARTICIPANTS

Fourteen right-handed, healthy participants (eight women, age 22–29, mean age 24.8) took part in the study. Each participant's laterality quotient, as assessed with the Edinburgh Handedness Inventory (Oldfield, 1971) was higher than 60. All participants were health screened by a physician and gave informed written consent.

### TASK-SYSTEMATIC

The movement repertoire consisted of five whole-body movements. Each movement consisted of three sub-movements which engaged a characteristic combination of extremities (**Figure 1**). Each of these movements was assigned an arbitrary name, comprising three syllables, each associated with one corresponding sub-movement (Ko-re-pa; Fe-so-da; Gu-la-mi; Ba-ki-te; Wa-ne-ro). None of these names is meaningful in German; neither were the combinations of the two first or last syllables of each. Importantly, in the course of the experiment, each movement (e.g., Ko-re-pa) could only be followed by one specific other (e.g., Fe-so-da) or by a repetition of itself (e.g., Ko-re-pa). Two piano chords, easily discernible even in absence of former musical training, were used to cue the transition between two movements. Each cue coincided with the onset of one movement and delivered an instruction on which movement was to *follow* the respective movement that had began when the cue sounded. The low chord meant that the transition following the current movement had to be a repetition (i.e., the same movement again; if the movement that started when the cue was presented was for example Ko-re-pa, the low chord meant it had to be followed by another Ko-re-pa). The high chord signaled that the transition following the current movement had to be a switch

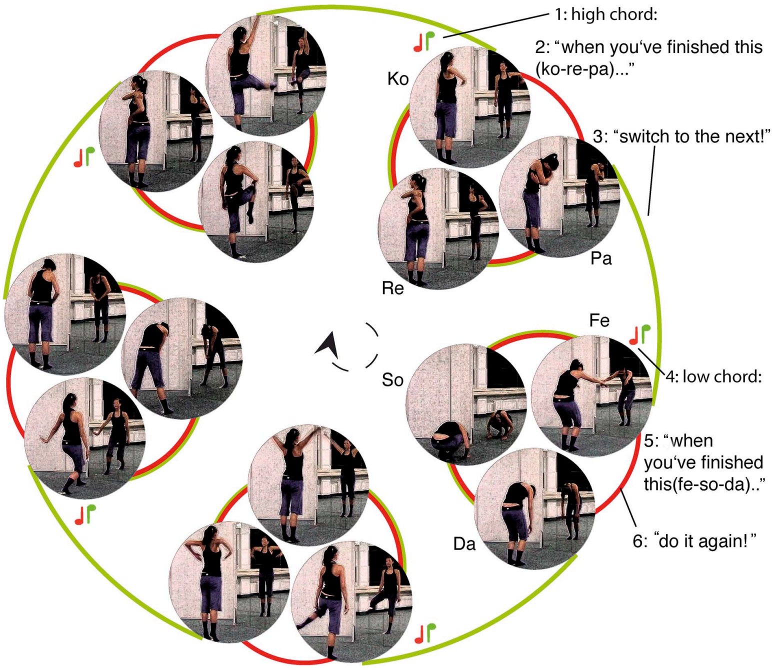

1: high chord:
2: "when you've finished this (ko-re-pa)..."
3: "switch to the next!"
Ko
Re
Pa
Fe
So
Da
4: low chord:
5: "when you've finished this(fe-so-da).."
6: "do it again!"

**FIGURE 1 | Movement sequence with depiction of respective sub-movements; the cue presented at each movement onset, determines the transition after completion of respective movement; low tune (red note): repetition of the movement (follows red line); high tune (green note): switch to the next movement (follow the green line).** Example: (i) a high chord sounds before ko-re-pa (2), indicating that the next movement must be fe-so-da (3 – a switch must take place); before fe-so-da starts, a low chord sounds (4) indicating that fe-so-da (5) must be followed by fe-so-da (6 repetition).

(i.e., the corresponding next movement; if the concurring movement was Ko-re-pa, the high chord meant it had to be followed by Fe-so-da). Switches were always switches to the next movement in a circular order (**Figure 1**), no movement was ever skipped. Thus, the upcoming movement was fully predictable, even if it differed from the current movement.

### SCHEDULE

The overall experimental schedule compromised three stages. Participants first had to pass a computer based behavioral experiment (stage 1) to be admitted to training (stage 2). If they completed training successfully, they were allowed to participate in the fMRI experiment (stage 3), which was virtually identical to the initial behavioral probe. The two test sessions and the movement training incorporated the same system of dynamically evolving movement sequences.

### STAGE 1: BEHAVIORAL PROBE SESSION

In the computer task, participants watched a dancer performing according to cue, but occasionally making mistakes. Previous to playing the task, the participants were instructed on the cue-movement associations that rule the task (low chord: repetition; high chord: switch). They received a short training where they could choose either four or eight example movies that contained up to 19 cued movements, before they started the task. The movie of the dancer was displayed in the middle of an otherwise gray computer screen, using the Software Presentation 12.0 (Neurobehavioral Systems, San Francisco, CA, USA). Visual input did not extend further than 5° of visual angle. The movies were stopped in irregular intervals and participants had to indicate by button press, whether the dancer had performed correctly immediately before video offset. That is, participants had to indicate whether the very last movement had been correct, irrespective of possible earlier errors. Questions were indicated by a question mark ("?") displayed in font size 24 for 1500 ms or until the first response. Participants had to press the arrow-to-the-left key (index finger) if they judged the last movement to have been according to cue or press the arrow-to-the-right key (middle finger) if they thought the movement had not been according to cue. Responses had to be given within a timeframe of 1500 ms and were followed by a valid feedback for 400 ms indicating correct, incorrect or delayed responses ("+"/"–"/"0"; **Figure 2**).

In both the behavioral and the fMRI session, the task encompassed 400 single movements. Thirty-two movements were not according to cue, i.e., the dancer switched to the next movement when a repetition had been cued (16), or a repetition was performed after a switch had been announced (16). Forty breaks disrupted the movie, which was thus divided into 41 videos of varying duration (3–17 movements each). In the behavioral experiment, all 40 breaks were question trials, 20 of them requesting an affirmative answer.

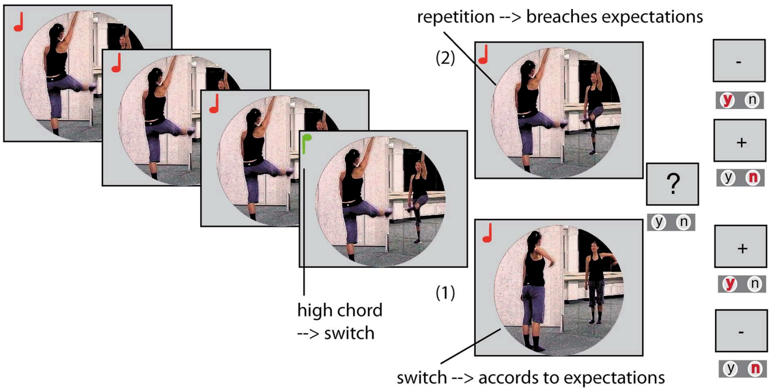

repetition --> breaches expectations
(2)
-
y n
+
y n
?
y n
+
y n
-
y n
(1)
high chord
--> switch
switch --> accords to expectations

**FIGURE 2 | Example for two alternative task-developments: Each picture is one movement within a movie, the notes signify high or low chords.** The actress could either perform the last movement before offset correctly (1), meaning the participants would have to judge the performance as correct by index finger button press ("y"). Or the last movement was not according to cue, a breach of expectation (2); the participants would then have to answer by middle-finger button press ("n") to judge the dancer's behavior as incorrect. Errors' during the movie where not to be judged upon. Movie offset and thus relevance of the error to the task was not predictable.

Up to four cues of the same kind, i.e., repetition or switch cues, could appear in a row. At the latest after the fourth identical transition cue (fourth switch or fourth repetition), a dissimilar transition was cued. Across the experiment, four, three, two, and one identical transitions after another appeared equally often. Thus, each cue that was dissimilar from the preceding cues could be differentiated from other first dissimilar cues by the number of preceding transitions that had been identical to each other. For example, a repetition cue could be the first repetition after two switches, or the first repetition after four switches. This randomization was employed to test for the assumptions of the saliency hypothesis (ii). The number of preceding different cues was a measurement of cue saliency against the backdrop of recent cue history.

### STAGE 2: MOVEMENT TRAINING

The participants that passed the 85% criterion of the behavioral experiment subsequently received six 1-h movement-training sessions within 10 days in order to establish a routine-like training stage for the cued performance of movement sequences. Training sessions were conducted in a small dance hall, one side walled with a mirror. During the first session, participants were taught the strict order of the five encompassed movements and learnt accurate performance of the single movements and the associated movement names. To that end, they were allowed to watch a model performing the respective movement on a laptop screen as often as they liked. Once the trainers were satisfied with accuracy of movement performance, participants were verbally instructed to conduct movement sequences, starting each movement when it was called out to them. In their second training session, participants learnt to move according to the cues. They started with a two-cue sequence. For example: Participants were told to start with the movement Ko-re-pa, as soon as the first cue sounded. If the cuing chord was low, they performed Ko-re-pa twice. Importantly they had to start the second Ko-re-pa after the second cue had rung. They had to withhold the third movement corresponding to the second cue, as they would only have been allowed to start the movement upon presentation of a third cue. Once every participant mastered this first step, the number of successive movements (cues) was constantly increased. If one or more participants made a mistake, the sequence had to be started from the beginning. This procedure was implemented in every training session forthwith, at the end of which participants mastered up to 18 cues in a row. Importantly, during training, more than four identical transitions in a row were possible. At the same time, participants had to keep moving at a high level of accuracy and trainers would correct them verbally, and, if necessary, by showing the model-video, over the entire course of training. At the end of the last training session, participants were filmed while performing three 15-cue sequences without further assistance (motor probe). During recording, they wore uniform clothing to allow for unbiased assessment of their performance in a later video evaluation.

### STAGE 3: fMRI SESSION

In the fMRI session that was scheduled for the day after each participant's respective last training session, they encountered the same task as in the behavioral probe. Participants lay supine on the scanner. Their head and arms were stabilized using form-fitting cushioning and their hands rested on a rubber foam tablet. On the right hand side, a response panel was mounted and fixed on the tablet. With their right hand index and middle finger resting on two response buttons, participants were able to judge on the correctness of the dancers movements within the same response contingencies as in the behavioral test. They wore earplugs to attenuate scanner noise and received auditory input via headphones. Participants received visual input on a mirror that was built into the head-coil and adjusted individually to allow for a comfortable view of the entire screen. All parameters were identical to the behavioral experiment (stage 1) with the exception that 24 breaks were used for question trials and 16 for null events (empty trials).

### DATA ACQUISITION

A 3T Siemens Magnetom Trio scanner (Siemens, Erlangen, Germany) was used in the functional imaging session. In a separate session, prior to the functional MRT, high-resolution 3D T1 weighted whole-brain MDEFT sequences were recorded for every participant (128 slices, field of view 256 mm, 256 × 256 pixel matrix, thickness 1 mm, spacing 0.25 mm). The functional session engaged a single-shot gradient echo-planar imaging (EPI) sequence sensitive to blood oxygen level dependent contrast (28 slices, parallel to the bicommisural plane, echo time 30 ms, flip angle 90°; repetition time 2000 ms; interleaved recording). Following the functional session immediately, a set of T1 weighted 2D-FLASH images was acquired for each participant (28 slices, field of view 200 mm, 128 × 128 pixel matrix, thickness 4 mm, spacing 0.6 mm, in-plane resolution 3 × 3 mm).

### fMRI DATA ANALYSIS

Functional data were offline motion-corrected using the Siemens motion protocol PACE (Siemens, Erlangen, Germany). Further processing was conducted with the LIPSIA (Lohmann et al., 2001) software package. Cubic-spline interpolation was used to correct for the temporal offset between the slices acquired in one scan. To remove low-frequency signal changes and baseline drifts, a 1/80 Hz filter was applied. The matching parameters (6 degrees of freedom, 3 rotational, 3 translational) of the T1 weighted 2-D FLASH data onto the individual 3-D MDEFT reference set were used to calculate the transformation matrice for linear registration. These Matrices were subsequently normalized to a standardized Talairach brain size ($x = 135$, $y = 175$, $z = 120$ mm; Talairach and Tournoux, 1988) by linear scaling. The normalized transformation matrices were then applied to the functional slices, to transform them using trilinear interpolation and align them with the 3-D reference set in the stereotactic coordinate system. The generated output had thus a spatial resolution of 3 mm × 3 mm × 3 mm (27 mm$^3$).

The statistical evaluation was based on a least-square estimation using the general linear model (GLM) for serially auto-correlated observations. Temporal Gaussian smoothing (4 s FWHM) was applied to deal with temporal autocorrelation and determine the degrees of freedom (Worsley and Friston, 1995). All design matrices were generated by hemodynamic modeling using a $\gamma$-function. We conducted the analysis once using only one GLM and once more using three GLMs (triple-GLM hereafter), one for each competing contrast. In the single-GLM approach, the three contrasts were set to compete for variance in one GLM to achieve a thorough model comparison. In the triple-GLM approach, the same whole-brain analyses was conducted with three separate GLMs, in order to not underestimate the effects of the competing alternative hypotheses and give them a more liberal chance to yield potential caudate activity. The onset vectors were modeled in a time-locked event-related fashion, i.e., the duration set to one second. The first derivative was taken into the model to improve model fit for latency effects.

### SINGLE-GLM APPROACH

The events to account for the violation hypothesis (i) were the dancer's incorrect movements, for the other hypotheses (ii and iii) the modeled events were specific cues. Hence, the model encompassed the following event types: correct movements, incorrect movements, first dissimilar transition cues, switch cues, and repetition cues (see below). The model encompassed null events as an additional vector. The violation hypothesis (i) contrasted invalidly cued switches and invalidly cued repetitions vs. validly cued switches and validly cued repetitions. The saliency hypothesis (ii) parametrically modeled the first dissimilar transition cue after 4, 3, 2, or 1 identical transition cue(s), ascribing the highest activation level (vector amplitude) to the dissimilar successor of four cues identical to each other. (The switch cue in **Table 2**, for example would have been assigned a vector amplitude of three; the immediately following repetition cue would have been assigned a vector amplitude of one). In addition, to discern whether potential effects would rely more on a first switch after a number of repetitions or first repetition after a number of switches, we modeled these contrasts in the same fashion of increasing vector values separately, too. That is, in one parametric contrast we ascribed a vector amplitude to each first switch according to the number of previous repetitions (saliency of switches contrast); in the other parametric contrast, the vector amplitude of each first repetition accorded to previous switches (saliency of repetitions contrast). The change hypothesis (iii) was modeled by comparing switch cues to repetition cues. Contrast images, i.e., differences of beta-value estimates for the specified conditions, were generated for each participant. All contrast images were fed into a second-level random effects analysis. The group analysis consisted of one-sample $t$-tests across all contrast images to analyze whether the observed differences between conditions were significantly deviant from 0. Acquired $t$-values were transformed to $z$-scores. To correct for false-positive results, an initial $z$-threshold was set to 2.56 ($p < 0.05$, one-tailed, uncorrected for multiple comparisons). In a second step, the results were corrected for multiple comparisons at the cluster level, using cluster size and cluster value thresholds that were obtained by Monte-Carlo simulations. The employed significance level was $p = 0.05$. Hence, the reported activations are significantly activated at $p \leq 0.05$, corrected for comparison at cluster level.

### TRIPLE-GLM APPROACH

In the triple-GLM approach the contrasts (i–iii) were calculated from the same events as described above. However, we employed a different GLM for each contrast that encompassed only the events necessary for the contrast and null-events. The GLM for the violation contrast (i) encompassed validly cued switches, validly cues repetition, invalidly cued switches and invalidly cues repetitions, and null-events. The GLM for the saliency contrast (i) encompassed all first dissimilar cues with a vector amplitude reflecting the number of previous identical repetitions and null-events. The GLM for the change hypothesis encompassed all repetition cues and all switch cues. Group analysis and corrections were identical to the single-GLM analysis described above.

## RESULTS

### BEHAVIORAL

Fifteen of 19 volunteers passed the initial behavioral probe at the 85% criterion. All participants completed training and responded correctly to cue – sequences of up to 18 cues. In the fMRI session, 14 out of 15 participants performed to criterion, with a mean rate of 91.1% correct responses, standard deviation (SD) at 5.4%. Mean rate of correct rejections was 91.7% (SD = 7.3%) and that of hits equaled 90.5% (SD = 7.2%). A two-tailed *t*-test revealed no significant difference between the averages of hits and correct rejections. One participant had to be excluded from further analyses due to insufficient performance (below 2 SDs from mean).

## fMRI

### SINGLE-GLM APPROACH

The contrast between movements that deviated from cue and movements that accorded to the previous cue (violation hypothesis, i) yielded significant bilateral activations in the basomedial caudate nucleus (**Figure 3**) and right medial pallidum, in the habenula, the anterior dorsal insula, mesial frontomedian cortex (Brodmann's area [BA] 8 and 9), lateral BA 10, and intraparietal sulcus (IPS). Significant lateralized activations were found in left angular gyrus (AG), right posterior superior temporal sulcus (pSTS) and right temporo-parietal junction (TPJ; **Figure 4**; **Table 1**).

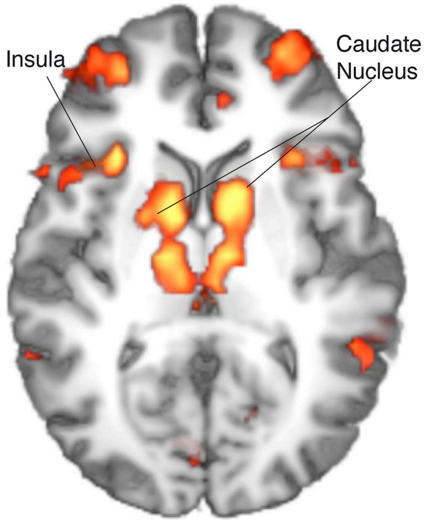


**FIGURE 3 | Single-GLM, violation hypothesis (i) contrast: main effect of movements deviating from cue vs. movements according to cue; group averaged activations are shown (at $z > 3.09$) on an axial slice ($z = 10$) of an individual brain, normalized and aligned to the Talairach stereotactic space.** Refer to **Table 1** for activation coordinates.

In the parametric contrasts testing for the saliency hypothesis (ii) only the contrast accounting for activity increase with the number of prior repetitions of identical movements revealed significant activation (saliency of switches contrast). This activity was in supplementary motor area (SMA) and postcentral gyrus. There were no significant correlations with the number of preceding switches. Similarly, there was no significant activation for the general saliency effect, that is number of identical transitions preceding a dissimilar transition, pooled over switches and repetitions. Contrasting switch cues with repetition cues to account for the change hypothesis (iii) yielded the right middle temporal gyrus (MTG) and bilaterally (anterior) IPS. Notably, there was no significant striatal activation, neither in the parametric contrasts relating to the saliency hypothesis (ii), nor in the contrast relating to change hypothesis (iii; **Table 2**).

### TRIPLE-GLM APPROACH

The triple-GLM analysis was employed to calculate the same contrasts as the single-GLM analysis but from three GLMs, optimized for differential effects. This approach yielded caudate activity also only in the violation contrast (i; **Table 3**; **Figure 5**). There was no striatal activity either in the saliency (ii) or change contrast (iii). Likewise, cortical activations identified by the violation (i; **Table 3**, **Figures 5 and 6**) contrast did not differ largely between the analogous contrasts from the single GLM vs. triple-GLM analyses. The triple-GLM analysis also revealed no significant activity for the saliency (ii) contrast. The parallel change (iii) contrasts from the two analysis approaches revealed quite similar patterns (**Tables 2 and 4** for comparison).

## DISCUSSION

The present study set out to investigate the role of the caudate nucleus in events that violate predictions (i). In contrast to previous studies, these events were not feedback in an operant conditioning task and involved neither reinforcement nor punishment. Hence, we termed these prediction-violating events "breaches of expectation" to distinguish them from prediction errors conceived as activity dependent on (future) reception of reward. Moreover, we extended the study-design to exclude the possibility that the striatal activity could be a consequence of potential secondary characteristics of violated predictions, that is responses to salient events (ii) and events that provoke a change in behavior (iii).

The contrast accounting for the violation hypothesis (i) yielded activation in the basomedial caudate nucleus. On the contrary, striatal activation was absent in the contrasts that accounted for

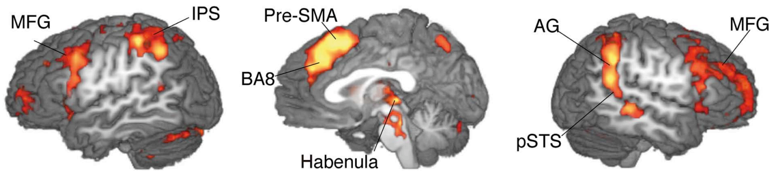


**FIGURE 4 | Single-GLM, violation hypothesis (i) contrast: main effect of movements deviating from cue vs. movements according to cue; group averaged activations are shown (at $z > 3.09$) on sagittal slices ($x = -52; 0; 52$) of an individual brain, normalized and aligned to the Talairach stereotactic space.** Refer to **Table 1** for activation coordinates.

**Table 1 | Single-GLM, violation hypothesis (i) contrast: Anatomical specification, Talairach coordinates (*x,y,z*) and maximal *z*-scores of significantly activated voxels for prediction-violating in contrast to prediction-conform movements.**

| Localization | Talairach coordinates | | | *z*-values, local maxima |
|---|---|---|---|---|
| | *x* | *y* | *z* | |
| Superior frontal gyrus (SFG)/pre-SMA (BA 8/6) | –2 | 21 | 45 | 5.8 |
| | 4 | 36 | 27 | 5.6 |
| Middle frontal gyrus (MFG; BA 8/9) | –41 | 15 | 39 | 5.1 |
| | 37 | 9 | 30 | 4.9 |
| | 34 | 33 | 30 | 5.3 |
| Dorsolateral prefrontal cortex (dlPFC), BA 10 | 31 | 54 | 18 | 4.7 |
| | –26 | 48 | 6 | 4.9 |
| Dorsal anterior insula | –32 | 21 | 0 | 5.9 |
| | 28 | 18 | 0 | 6.2 |
| Angular gyrus (AG) | –56 | –45 | 36 | 4.9 |
| Inferior parietal lobule (LPI) | 34 | –51 | 45 | 5.3 |
| | –53 | –51 | 39 | 4.8 |
| Intraparietal sulcus (IPS) | –41 | –36 | 39 | 5.6 |
| | 52 | –45 | 33 | 5.1 |
| Posterior cingulate cortex (BA 23) | –5 | –21 | 30 | 3.5 |
| | 7 | –33 | 30 | 3.9 |
| Posterior superior temporal sulcus (pSTS) | 49 | –36 | 0 | 4.5 |
| Temporal-parietal junction (TPJ) | –50 | –48 | 12 | 3.9 |
| Precuneus | –8 | –66 | 45 | 5.1 |
| Basomedial head of caudate nucleus (CAU) | –11 | 6 | 6 | 5.6 |
| | 10 | 9 | 9 | 5.4 |
| Medial globus pallidus (GPi) | 13 | 0 | 3 | 5.7 |
| Habenula | 1 | –27 | 3 | 5.7 |
| Thalamus, ventrolateral nucleus (VL) | –14 | –15 | 3 | 5.4 |
| | –14 | –12 | 12 | 4.7 |
| Nucleus ruber | –8 | –27 | –6 | 4.9 |
| | 4 | –27 | –6 | 5.1 |
| Cerebellum | 16 | –48 | –15 | 4.1 |
| | –17 | –84 | –21 | 4.9 |
| | 28 | –54 | –24 | 4.6 |
| | –32 | –60 | –24 | 5.0 |

the saliency (ii) or change (iii) hypotheses, even when we calculated these contrasts from separate optimized GLMs in the triple-GLM approach. This pattern of results suggests the dorsal striatum to be tuned to violations of current predictions rather than to these events' saliency or implied incite to switch behavior. Moreover, the results show that dorso-striatal responses to violated predictions are not restricted to reinforcement or punishment protocols.

### CAUDATE NUCLEUS SIGNALS FOR BREACHES OF EXPECTATION

The results of the current study suggest that activity in the head of caudate nucleus signals breaches of expectation, i.e., violated predictions, more generally than previously assumed. This finding may explain why this area is often found in trial and error learning, where its activity diminishes once learning has occurred (Jueptner and Weiller, 1998; Delgado et al., 2005; Shohamy et al., 2008; Ruge and Wolfensteller, 2009). Trial and error learning means building up predictions what cues demand which action to gain reward. If the predictions fail, actions have to be altered. If the predictions are fulfilled, the deterministics of the task are apparently understood. Accordingly, the diminution of caudate activity over the course of learning is rooted in the fact that only as long as the rules of the task are unknown (at the beginning of learning), predictions are constantly violated, driving caudate activity. Once the rules have been established, breaches of expectation wane and so does caudate activity. The notion that a breach of expectation signal is generated in caudate nucleus could also account for impairments of Parkinson's disease patients in trial and error learning (Shohamy et al., 2008). Their compromised breach of expectation signal, due to neostriatal dysfunction, hinders updating wrong beliefs and accordingly adapting behavior. More evidence for a caudate signal for breaches of expectation comes from studies showing that caudate activity ceases the earlier, the easier it is to learn the association between cues and correct actions, i.e., the easier it is to build up operative predictions (Delgado et al., 2005; Koch et al., 2008). The same

**Table 2 | Single-GLM, change hypothesis (iii) contrast: Anatomical specification, Talairach coordinates (*x,y,z)* and maximal *z*-scores of significantly activated voxels for prediction-violating in contrast to prediction-conform movements.**

| Localization | Talairach coordinates | | | *z*-values, local maxima |
|---|---|---|---|---|
| | *x* | *y* | *z* | |
| Dorsal premotor cortex (PMd) | 28 | 0 | 54 | 3.6 |
| | –26 | 6 | 60 | 5.4 |
| Middle frontal gyrus | 31 | 42 | 24 | 4.1 |
| Presupplementary motor area (pre-SMA) | –5 | 9 | 48 | 3.6 |
| Inferior frontal junction (IFJ) | –35 | 6 | 30 | 4.1 |
| Superior parietal lobule (SPL) | 19 | –54 | 60 | 5.0 |
| | –14 | –69 | 48 | 5.0 |
| Intraparietal sulcus (IPS) | –38 | –42 | 51 | 4.2 |
| | –29 | –75 | 30 | 4.1 |
| Posterior middle temporal gyrus (pMTG) | 43 | –69 | 3 | 4.0 |
| | –53 | –66 | 6 | 4.2 |
| | –47 | –51 | –9 | 5.4 |
| Cuneus | –20 | –96 | 3 | 4.0 |
| Thalamus | –14 | –15 | 12 | 4.6 |
| Cerebellum | 10 | –51 | –36 | 4.2 |

activity is persistent for cues that are non-informative, and this is for the same reasons, i.e., that they predict that either of two actions could be correct with the same probability. These cues make it impossible to establish reliable predictions (Delgado et al., 2005).

The present findings add to these results in an important fashion, showing caudate nucleus involvement for general breaches of expectation, independent of ensuing feedback, in a movement observation paradigm. Breaches of expectation yielded caudate activity, even if the violated prediction was not a prediction on the availability of reinforcement, but only on the next movement that was to be observed. This finding of a "perceptual prediction error" (Zacks et al., 2007) stands in stark contrast to the aforementioned studies that investigated the caudate prediction error signal in relation to feedback on whether the participants had gained or lost money by their last action (Delgado et al., 2005; Koch et al., 2008; Tricomi and Fiez, 2008) Moreover, as the breaches of expectation in this study reflect perceptual prediction error, it establishes that this perceptual prediction error has a neural correlate in caudate nucleus.

### CORTICAL AREAS CO-ACTIVE WITH CAUDATE NUCLEUS

We found a number of cortical areas co-activated for the violation contrast, including the right posterior superior temporal sulcus (pSTS) and the adjacent tempo-parietal junction (TPJ) extending into AG. All three cortical regions are connected to the neostriatum by the fronto-occipital fasciculus as well as by the joint fasciculus subcallosal of Muratoff (Schmahmann and Pandya, 2006). This white-matter connectivity points to functionally closely interrelated areas. As the Muratoff bundle directly projects into the dorsal striatum, a fast transmission of perceived deviations to neostriatum is accounted for. Moreover, projections from the AG to caudate nucleus have recently been confirmed by diffusion-tensor imaging in humans (Uddin et al., 2010).

With regard to this connectivity and these regions' functions as described in the literature, the concurrent activation for the violation contrast is quite plausible. The pSTS is activated when perceiving biological motion and shows enhanced activation for movements that deviate from expectations (Keysers and Perrett, 2004). Adjacent TPJ is involved in predicting the end-state of movements (Arzy et al., 2006) and also in reorienting in space (Blanke et al., 2004; Van Overwalle and Baetens, 2009). Accordingly, pSTS enhancement may indicate the dissimilarity between a covert motor plan in our highly trained subjects and the actually perceived (false) movement. TPJ activation, moreover, may result from perceiving limb trajectories toward end-states that differed from the expected (or even covertly prepared) ones. The spreading of activation into AG fosters the idea put forward by other authors that TPJ, extending into AG, actually responds to breaches of expectation (Vossel et al., 2006; Shulman et al., 2009). These authors employed paradigms where a number of cues signaled where a target would appear with different probabilities. Interestingly, when a cue that indicated a high probability of a certain target position was violated, the resulting activation was higher than for the violation of less predictive cues. It therefore seems as if the more surprising an outcome is, that is, the more it violates a current prediction, the higher is the resultant AG activation (Vossel et al., 2006; Shulman et al., 2009). Besides, the aforementioned studies used abstract stimuli, but employed paradigms demanding reorienting in space (Blanke et al., 2004; Vossel et al., 2006; Shulman et al., 2009; Van Overwalle and Baetens, 2009). The current study employed solely stimuli that represented human motion, but results agree with the literature that the function of reorienting of attention is related to activity in the temporo-parietal junction and posterior parietal cortex (Corbetta and Shulman, 2002). Within this framework, TPJ would be reframed as "circuit breaker," which still implies the same function of detecting a salient stimulus that deviates from expectations (Corbetta and Shulman, 2002).

It can be suggested that after transmission of the perceived violation of prediction from the temporo-parietal network to the dorsal striatum, a breach of expectation signal is provided by the dorsal striatum that incites the mediation of frontal responses (Ridderinkhof et al., 2004). The consequential frontal network comprised the mesial frontal cortex bilaterally, specifically Brodmann area 8 (BA), the anterior cingulate cortex (ACC) and the presupplementary motor area (pre-SMA), anterior dorsal insula and middle frontal gyrus (MFG). Of those, the anterior dorsal insula, ACC and mesial BA 8 have been implicated in situations characterized by uncertainty (Volz et al., 2003; Wager and Feldman, 2004; Volz, 2005), signifying the likelihood of errors and a need to adapt one's expectations. In the present study, participants expected to encounter events that would deviate from current predictions, but they lacked information regarding the time-point and frequency of such deviating events. Consequently, there was uncertainty toward the ruling probabilistic of the task.

**Table 3 | Triple-GLM, violation hypothesis (i) contrast: Anatomical specification, Talairach coordinates (*x,y,z*) and maximal *z*-scores of significantly activated voxels for prediction-violating in contrast to prediction-conform movements.**

| Localization | Talairach coordinates | | | *z*-values, local maxima |
|---|---|---|---|---|
| | *x* | *y* | *z* | |
| Superior frontal gyrus (SFG)/pre-SMA (BA 8/6) | –2 | 21 | 48 | 5.8 |
| | 4 | 36 | 27 | 5.6 |
| Middle frontal gyrus (MFG; BA 8/9) | –41 | 15 | 39 | 5.1 |
| | 37 | 15 | 36 | 4.8 |
| | 37 | 33 | 27 | 5.0 |
| Dorsolateral prefrontal cortex (dlPFC), BA 10 | 31 | 54 | 18 | 4.8 |
| | –26 | 48 | 6 | 4.9 |
| Dorsal anterior insula | –32 | 21 | 3 | 5.8 |
| | 28 | 18 | –3 | 6.0 |
| Angular Gyrus (AG) | –56 | –45 | 36 | 4.9 |
| Inferior parietal lobule (LPI) | 34 | –51 | 45 | 5.3 |
| | –53 | –51 | 39 | 4.8 |
| Intraparietal sulcus (IPS) | –41 | –36 | 39 | 5.6 |
| | 52 | –45 | 33 | 5.1 |
| Posterior cingulate cortex (BA 23) | –2 | –21 | 33 | 3.6 |
| | 7 | –33 | 30 | 3.7 |
| Posterior superior temporal sulcus (pSTS) | 49 | –36 | 0 | 4.6 |
| Temporal-parietal junction (TPJ) | –50 | –48 | 12 | 3.6 |
| Precuneus | –5 | –66 | 42 | 4.7 |
| Basomedial head of caudate nucleus (CAU) | –11 | 6 | 6 | 5.7 |
| | 10 | 9 | 9 | 5.4 |
| Medial globus pallidus (GPi) | 13 | 0 | 3 | 5.8 |
| Habenula | 1 | –27 | 3 | 5.4 |
| Thalamus, ventrolateral nucleus (VL) | –14 | –15 | 3 | 5.0 |
| | –14 | –12 | 12 | 4.7 |
| Nucleus ruber | –8 | –27 | –6 | 4.8 |
| | 4 | –27 | –6 | 5.0 |
| Cerebellum | 16 | –48 | –15 | 4.0 |
| | –17 | –81 | –21 | 4.9 |
| | 28 | –54 | –24 | 4.4 |
| | –32 | –60 | –24 | 5.0 |

The involvement of MFG signifies the high impact the task had on working memory (Goldman-Rakic, 1987; Braver et al., 1997). The participants had to judge a movement according to an auditory cue that had preceded the currently presented cue. At the same time they had to register the current cue to predict the next movement. This protocol amounts to a one-back-task. Thus, MFG may reflect active retrieval of the last cue in order to judge whether the caudate-conveyed deviation signal was meaningful or not. Engaging working memory in response to a signaled deviation accords to the assumption that in uncertain situations, PFC explores alternatively operating models (Daw et al., 2005).

### SUBCORTICAL BREACH OF EXPECTATION CODING

Apart from caudate nucleus, an important subcortical component of the activity in the violation contrast was the habenula. The habenula codes exclusively for negative prediction errors, and thus corresponds in an antagonistic fashion to the classic reward prediction error (Hikosaka et al., 2008). Unexpected positive reinforcement causes a decrease in habenula activity. Meanwhile, during punishment or reward omission the habenula shows an activity increase (Matsumoto and Hikosaka, 2009). In contrast, caudate activity has been suggested to be different from both opponents in that it is increased in activity for every breach of expectation, regardless its valence (Horvitz, 2000). In the present study, the violations of current predictions did not entail negative consequences. However, making a mistake during behavioral training could well have been ascribed a negative valence. We suggest that as the participants probably engaged in motor imagery to solve the task, exploiting their own memorized experiences during training, seeing the dancer deviating from the protocol was regarded an error with all negative implications (Preston and de Waal, 2001; Decety and Jackson, 2004; Singer et al., 2004). However, the fact that in this study the signal to violated predictions in caudate nucleus is enhanced at the same time that the habenula codes for a negative prediction error, underpins the breach of expectation nature of caudate activity. We find an activity increase for events that are not predicted in the current forward model, not a typical negative prediction error, as the corresponding habenula activity may suggest (Jocham and Ullsperger, 2009).

### PREDICTIONS, DEVIATIONS, AND LEARNING

In the animal literature, prediction errors are mostly described as resultant from the occurrence or omission of reward (Schultz et al., 1997), thereby related to satisfying or averse (external) stimuli. Prediction errors are defined as decreased activity in the face of omitted reward or punishment. However, the current study revealed evidence for heightened caudate activity toward violated predictions in insufficient or failing forward models, and hence breaches of expectation. Establishing predictions and especially signaling for a breach of expectation may be as important as coding how much reward (e.g., food or money) is available, or how unpredicted primary reward was (Spicer et al., 2007). The limitations of fMRI in proving neurotransmitter involvement do not allow drawing the inference that this caudate activity was based in a dopaminergic response (Düzel et al., 2009). The dopaminergic innervation of the dorsal striatum (see Joel and Weiner, 2000 #197 for a review) and the response of the habenula (Jocham and Ullsperger, 2009) implicate the dopaminergic system, but further studies are needed to decide whether dopamine is involved in not reward related breaches of expectation. New approaches, especially the free-energy principle (Friston, 2010) stress the value of predictive capability *per se*, i.e., the ability to detect breaches of expectation (Kiebel et al., 2008; Suddendorf et al., 2009; Friston, 2010). This model regards correct predictions as prerequisite for

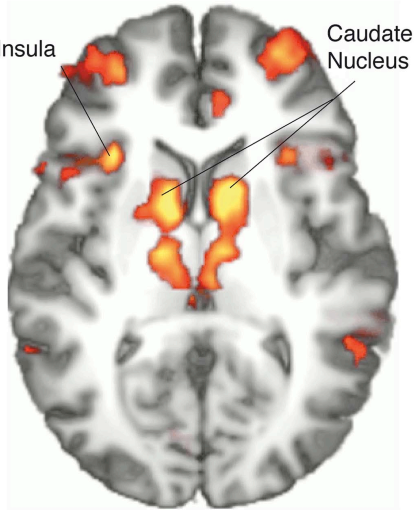


**FIGURE 5 | Triple-GLM, violation hypothesis (i) contrast: main effect of movements deviating from cue vs. movements according to cue; group averaged activations are shown (at $z > 3.09$) on an axial slice ($z = 10$) of an individual brain, normalized and aligned to the Talairach stereotactic space.** Refer to **Table 3** for activation coordinates.

**Table 4 | Triple-GLM, change hypothesis (iii) contrast: Anatomical specification, Talairach coordinates (*x,y,z*) and maximal *z*-scores of significantly activated voxels for prediction-violating in contrast to prediction-conform movements.**

| Localization | Talairach coordinates | | | *z*-values, local maxima |
|---|---|---|---|---|
| | *x* | *y* | *z* | |
| Dorsal premotor cortex (PMd) | –20 | –6 | 51 | 4.4 |
| Ventral premotor cortex (PMv) | –53 | 6 | 33 | 4.1 |
| Middle frontal gyrus (MFG) | 43 | 24 | 39 | 3.9 |
| Superior frontal gyrus (SFG), BA 8 | –8 | 27 | 36 | 3.7 |
| Presupplementary motor area (pre-SMA) | –5 | 9 | 48 | 4.8 |
| Superior parietal lobule (SPL) | 22 | –57 | 63 | 4.0 |
| | –26 | –51 | 57 | 3.9 |
| Intraparietal sulcus (IPS) | –50 | –27 | 33 | 4.4 |
| | 34 | –30 | 42 | 4.0 |
| | 25 | –63 | 42 | 3.9 |
| Posterior middle temporal gyrus (pMTG) | 40 | –57 | 6 | 4.8 |
| | –53 | –69 | 3 | 5.1 |
| | –44 | –51 | –9 | 3.7 |
| Precuneus | 7 | –51 | 57 | 3.9 |
| Lingual gyrus | 13 | –96 | –12 | 3.6 |

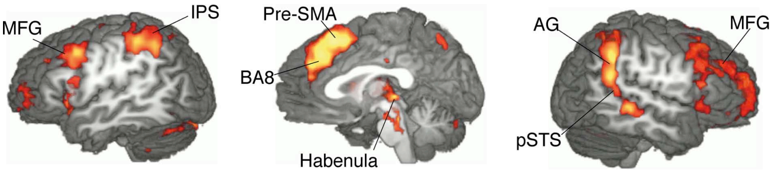


**FIGURE 6 | Triple-GLM, violation hypothesis (i) contrast: main effect of movements deviating from cue vs.movements according to cue; group averaged activations are shown (at $z > 3.09$) on sagittal slices ($z = –52; 0; 52$) of an individual brain, normalized and aligned to the Talairach stereotactic space.** Refer to **Table 3** for activation coordinates.

successful interaction with the environment, because recognizing a situation and acting accordingly is a capability owed to the disposal of valid forward models (Kiebel et al., 2008). The according actions may be to the end of satisfying primary needs or an evolved want. A typical aim in a social interaction would be, e.g., to adhere to the arrangement and to the task instruction in an fMRI study that were formerly mutually agreed upon. Operative forward models themselves can be valuable enough to be perceived as rewarding, even though they do not yield primary reward or reinforcement. Consider, for example, the psychological importance of a sense of control. In learned-helplessness paradigms, where animals are not able to predict and avoid punishment, pseudo-depression is a consequence (Seligman and Maier, 1967). Unpredictability is just another facet of non-operative forward-models. To establish operative forward models, breaches of expectation must be registered. If they cease to occur, this can be regarded as evidence that learning, i.e., model adaptation, was sufficient. The sense of accomplishment that goes with feeling in control of a situation provides indeed a powerful motivation to learn. Taken together, breaches of expectation that allow for the generation and improvement of an internal model are of utmost importance to survival, but also psychological wellbeing.

## CONCLUSION

The results of the current study foster the idea that the caudate nucleus signals for occurrence of events that violate the predictions of the operative forward model. This signal is not due to the perception of salient events or the need to change one's behavior, and it is not based on direct reinforcement or punishment. Frontal activation that we observed may be triggered by this signal from the caudate nucleus and operate to deal with present altered environmental demands; either via update of the current forward model or via assessment of the probability of certain event alternatives.

## SUPPLEMENTARY MATERIAL

The Movies S1 and S2 for this article can be found online at http://www.frontiersin.org/human_neuroscience/10.3389/fnhum.2011.00038/abstract/

## REFERENCES

Alexander, G. E., DeLong, M. R., and Strick. P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.* 9, 357–381.

Arzy, S., Thut, G., Mohr, C., Michel, C. M., and Blanke, O. (2006). Neural basis of embodiment: distinct contributions of temporoparietal junction and extrastriate body area. *J. Neurosci.* 26, 8074–8081.

Badgaiyan, R. D., Fischman, A. J., and Alpert, N. M. (2007). Striatal dopamine release in sequential learning. *Neuroimage* 38, 549–556.

Blanke, O., Theodor, L., Laurent, S., and Margitta, S. (2004). Out-of-body experience and autoscopy of neurological origin. *Brain* 127, 243–258.

Braver, T. S., Cohen, J. D. Nystrom, L. E. Jonides, J. Smith, E. E., and Noll. D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage* 5, 49–62.

Bubic, A., Yves Von Cramon, D., and Schubotz, R. I. (2009). Prediction, cognition and the brain. *Front. Hum. Neurosci.* 5:12. doi: 10.3389/fnhum.2010.00025

Bunge, S. A., Kahn, I., Wallis, J. D., Miller, E. K., and Wagner, A. D. (2003). Neural circuits subserving the retrieval and maintenance of abstract rules. *J. Neurophysiol.* 90, 3419–3428.

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215.

Davidson, M. C., Horvitz, J. C., Tottenham, N., Fossella, J. A., Watts, R., Ulug, A. M., and Casey, B. J. (2004). Differential cingulate and caudate activation following unexpected nonrewarding stimuli. *Neuroimage* 23, 1039–1045.

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

Decety, J., and Jackson, P. L. (2004). The functional architecture of human empathy. *Behav. Cogn. Neurosci. Rev.* 3, 71–100.

Delgado, M. R., Miller, M. M., Inati, S., and Phelps, E. A. (2005). An fMRI study of reward-related probability learning. *Neuroimage* 24, 862–873.

den Ouden, H. E., Danizeau, J., Roiser, J., Friston, K. J., and Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *J. Neurosci.* 30, 3210–3219.

den Ouden, H. E., Friston, K. J., Daw, N. D., McIntosh, A. R., and Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebr. Cortex* 19, 1175.

Düzel, E., Bunzeck, N., Guitart-Masip, M., Wittmann, B., Schott, B. H., and Tobler, P. N. (2009). Functional imaging of the human dopaminergic midbrain. *Trends Neurosci.* 32, 321–328.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138.

Goldman-Rakic, P. S. (1987). "Circuitry of primate prefrontal cortex and regulation of behavior by representational memory," in *Handbook of Physiology – The Nervous System,* Vol. 5. eds F. Plum and V. Mountcastle (Bethesda), 373–417.

Grahn, J. A., Parkinson, J. A., and Owen, A. M. (2008). The cognitive functions of the caudate nucleus. *Prog. Neurobiol.* 86, 141–155.

Graybiel, A. M. (2005). The basal ganglia: learning new tricks and loving it. *Curr. Opin. Neurobiol.* 15, 638–644.

Hikosaka, O., Sesack, S. R., Lecourtier, L., and Shepard, P. D. (2008). Habenula: crossroad between the basal ganglia and the limbic system. *J. Neurosci.* 28, 11825–11829.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience* 96, 651–656.

Jocham, G., and Ullsperger, M. (2009). Neuropharmacology of performance monitoring. *Neurosci. Biobehav. Rev.* 33, 48–60.

Joel, D., and Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, 96, 451–474.

Jueptner, M., and Weiller, C. (1998). A review of differences between basal ganglia and cerebellar control of movements as revealed by functional imaging studies. *Brain* 121, 1437–1449.

Keysers, C., and Perrett, D. I. (2004). Demystifying social cognition: a Hebbian perspective. *Trends Cogn. Sci.* 8, 501–507.

Kiebel, S. J., Daunizeau, J., and Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* 4:e1000209. doi: 10.1371/journal.pcbi.1000209

Kilner, J. M., Friston, K. J., and Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cogn. Process* 8, 159–166.

Koch, K., Claudia, S., Gerd, W., Reichenbach, J. R., Sauer, H., and Schlösser, R. (2008). The neural correlates of reward-related trial-and-error learning: an fMRI study with a probabilistic learning task. *Learn. Mem.* 15, 728–732.

Lohmann, G., Karsten, M., Volker, B., Heiko, M., Sven, H., Lin, C., Zysset, S., and Yves von Cramon, D. (2001). Lipsia – a new software system for the evaluation of functional magnetic resonance images of the human brain. *Comput. Med. Imaging Graph* 25, 449–457.

Matsumoto, M., and Hikosaka, O. (2009). Representation of negative motivational value in the primate lateral habenula. *Nat. Neurosci.* 12, 77–84.

O'Doherty, J. P., Buchanan, T. W., Seymour, B., and Dolan, R. J. (2006). Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron* 49, 157–166.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9, 97–113.

Preston, S. D., and de Waal, F. B. (2001). Empathy: its ultimate and proximate bases. *Behav. Brain Sci.* 25, 1–71.

Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975.

Rescorla, R. A., and Wagner, A. W. (1972). "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory,* eds A. H. Black and W. F. Prokasy (New York: Appleton-Century-Crofts), 64–99.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., and Nieuwenhuis, S.

(2004). The role of the medial frontal cortex in cognitive control. *Science* 306, 443–447.

Roy, E. A., Saint-Cyr, J., Taylor, A., and Lang, A. (1993). Movement sequencing disorders in Parkinson's disease. *Int. J. Neurosci.* 73, 183–194.

Ruge, H., and Wolfensteller, U. (2009). Rapid formation of pragmatic rule representations in the human brain during instruction-based learning. *Cereb. Cortex* 20, 1656–1667.

Saint-Cyr, J. A. (2003). Frontal-striatal circuit functions: context, sequence, and consequence. *J. Int. Neuropsychol. Soc.* 9, 103–127.

Schmahmann, J. D., and Pandya, D. N. (2006). *Fiber Pathways of The Brain,* 1 Edn. New York: Oxford University Press.

Schultz, W. (2000). Multiple reward signals in the brain. *Nat. Rev. Neurosci.* 1, 199–207.

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

Schultz, W., and Dickinson, A. (2000). Neuronal coding of prediction errors. *Ann. Rev. Neurosci.* 23, 473–500.

Schultz, W., Tremblay, L., and Hollerman, J. R. (1998). Reward prediction in primate basal ganglia and frontal cortex. *Neuropharmacology* 37, 421–429.

Schutz-Bosbach, S., and Prinz, W. (2007). Prospective coding in event representation. *Cogn. Process* 8, 93–102.

Seligman, M. E., and Maier, S. F. (1967). Failure to escape traumatic shock. *J. Exp. Psychol.* 74, 1–9.

Shohamy, D., Myers, C. E., Kalanithi, J., and Gluck, M. A. (2008). Basal ganglia and dopamine contributions to probabilistic category learning. *Neurosci. Biobehav. Rev.* 32, 219–236.

Shulman, G. L., Astafiev, S. V., Franke, D., Pope, D. L., Snyder, A. Z., McAvoy, M. P., and Corbetta, M. (2009). Interaction of stimulus-driven reorienting and expectation in ventral and dorsal frontoparietal and basal ganglia-cortical networks. *J. Neurosci.* 29, 4392–4407.

Singer, T., Seymour, B. O'Doherty, J. P., Kaube, H., Dolan, R., J., and Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain *Science* 303, 1157–1162.

Spicer, J., Galvan, A., G., Hare, T. A., Voss, H. Glover, G., and Casey, B. J. (2007). Sensitivity of the nucleus accumbens to violations in expectation of reward. *Neuroimage* 34, 455–461.

Suddendorf, T., Donna, R. A., and Corballis, M. C. (2009). Mental time travel and the shaping of the human mind. *Philos. Trans. R. Soc. B. Biol. Sci.* 364, 1317–1324.

Talairach, J., and Tournoux, P. (1988). Co-planar stereotaxic atlas of the human brain. New York: Thieme.

Tricomi, E., and Fiez, J. A. (2008). Feedback signals in the caudate reflect goal achievement on a declarative memory task. *Neuroimage* 41, 1154–1167.

Uddin, L. Q., Supekar, K., Amin, H., Rykhlevskaia, E., Nguyen, D. A., Greicius, M. D., and Menon, V. (2010). Dissociable connectivity within human angular gyrus and intraparietal sulcus: evidence from functional and structural connectivity. *Cereb. Cortex* 20, 2636–2646.

Van Overwalle, F., and Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage* 48, 564–584.

Volz, K. G. (2005). Variants of uncertainty in decision-making and their neural correlates. *Brain Res. Bull.* 67, 403–412.

Volz, K. G., Schubotz, R. I., and von Cramon, D. Y. (2003). Predicting events of varying probability: uncertainty investigated by fMRI. *Neuroimage* 19(2 Pt 1), 271–280.

Vossel, S., Thiel, C. M., and Fink, G. R. (2006). Cue validity modulates the neural correlates of covert endogenous orienting of attention in parietal and frontal cortex. *Neuroimage* 32, 1257–1264.

Wager, T. D., and Feldman, B. L. (2004). From affect to control: functional specialization of the insula in motivation and regulation. Retrieved from www.apa.org/psycextra/ on 29 March 2011.

Wolpert, D. M., and Flanagan, J. R. (2001). Motor prediction. *Curr. Biol.* 11, 729–732.

Worsley, K. J., and Friston, K. J. (1995). Analysis of FMRI time series revisited – again. *Neuroimage* 2, 173–181.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychol. Bull.* 133, 273–293.

Zink, C. F., Pagnoni, G., Martin, M. E., Dhamala, M., and Berns, G. S. (2003). Human striatal response to salient nonrewarding stimuli. *J. Neurosci.* 23, 7.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 14 January 2011; accepted: 27 March 2011; published online: 08 April 2011.*

*Citation: Schiffer A-M and Schubotz RI (2011) Caudate nucleus signals for breaches of expectation in a movement observation paradigm. Front. Hum. Neurosci.* ***5****:38. doi: 10.3389/fnhum.2011.00038*

## 2.2 Neural Changes When Actions Change: Adapatation of Strong and Weak Expectations

Bias article

◆ Human Brain Mapping 000:000–000 (2011) ◆

# Neural Changes When Actions Change: Adaptation of Strong and Weak Expectations

Anne-Marike Schiffer,[1]* Christiane Ahlheim,[1]
AQ1 Kirstin Ulrichs,[1] and Ricarda I. Schubotz[1,2]

[1]Motor Cognition Group, Max Planck Institute for Neurological Research, Cologne, Germany
AQ2 [2]Westfälische Wilhelms-Universität Münster, Institut für Psychologie, Münster, Germany

**Abstract:** Repeated experiences with an event create the expectation that subsequent events will expose an analog structure. These spontaneous expectations rely on an internal model of the event that results from learning. But what happens when events change? Do experience-based internal models get adapted instantaneously, or is model adaptation a function of the solidity of, i.e., familiarity with, the corresponding internal model? The present fMRI study investigated the effects of model solidity on model adaptation in an action observation paradigm. Subjects were made acquainted with a set of action movies that displayed an altered script when encountered again in the scanning session. We found model adaptation to result in an attenuation of the premotor-parietal network for action observation. Model solidity was found to modulate activation in the parahippocampal gyrus and the anterior cerebellar lobules, where increased solidity correlated with activity increase. Finally, the comparison between early and late stages of learning indicated an effect of model solidity on adaptation rate. This contrast revealed the involvement of a fronto-mesial network of Brodmann area 10 and the ACC in those states of learning that were signified by high model solidity, no matter if the memorized original or to the altered action model was the more solid component. Findings suggest that the revision of an internal model is dependent on its familiarity. Unwarranted adaptations, but also perseverations may thus be prevented. *Hum Brain Mapp 00:000–000, 2011.* © 2011 Wiley Periodicals, Inc.

**Key words:** forward model; frontal pole; action observation; adaptation; breach of expectation; fMRI

## INTRODUCTION

We don’t inspect events without expecting their course. According to the predictive coding account of action observation, action perception triggers an “internal model” [Kilner et al., 2007; Neal and Kilner, 2010] that is run in real time and consists of predictions on the course of action [Schutz-Bosbach and Prinz, 2007]. Evidently, such predictions save resources [Zacks et al., 2007].

However, it is not only of tremendous importance to establish internal models through experience, but also to attune them to persistent changes, and thus maintain valid predictions. Consider being forced to change your well-known way to work because of some indiscernible traffic condition at some point of the route. If this happens once, you may surely assume that something like a traffic accident has happened. In all probability you would not decide to take another way to work on the next day. This is an example of a well-established and therefore solid internal model being violated. Solidity means that a model has strong connection weights between encompassed events. Events that have through repeated exposure become very well associated with each other elicit implicit

*Correspondence to: Anne-Marike Schiffer, Motor Cognition Group, Max Planck Institute for Neurological Research, Gleueler Straße 50, 50931 Cologne, Germany. E-mail: schiffer@nf.mpg.de

Received for publication 12 April 2011; Revised 13 October 2011; Accepted 28 November 2011

DOI: 10.1002/hbm.22023

Published online in Wiley Online Library (wileyonlinelibrary.com).

prediction of each other. Solidity, i.e., a large strength of association, determines that the deviation is treated as a one-time occurrence of no further importance for future predictions.

Now consider being on holiday and the road to the beach being blocked on your second day in the unfamiliar countryside. You may start wondering whether you have chosen exactly the way you went the day before and try to reverse the mental map you have created of your surroundings. If you find a new way to the beach and follow it on all occasions thereafter, you may quite forget, or begin to doubt that another way has ever been possible. This form of adaptation seems likely in case of low familiarity, i.e., a weak internal model. The weak internal model is questioned and possibly revised after a one-time breach of expectation. However, it remains to be experimentally established how an internal model's solidity influences its revision and hence adaptation of predictions. To our knowledge only a few studies on reversal learning in stimulus–response paradigms [Ghahremani et al., 2009] have dealt with the influence of model solidity on adaptation; no study has addressed the question in an action observation paradigm.

The present fMRI study was designed to investigate the influence of model solidity on its adaptation during iterations of a divergent script. Internal models of different solidity were established by presenting a number of scripts, i.e., movies showing everyday actions (as will be described below in more detail). The concept of solidity is similar to associative strength [McClelland et al., 1995] between components of an internal representation. Thus, solidity pertains to an internal model whose constituent events are highly associated with each other. Hence, in a fixed temporal schedule, each constituent elicits prediction of the next. This prediction is a consequence of statistical learning [Turke-Brown et al., 2010]. Statistical learning results from repeated pairing of events, i.e., stimulus familiarity, that has been proposed to be critical in extending the persistence of memory [Eichenbaum, 2000]. Concisely, repeated exposure leads to solidity. In a solid model, each event is highly associated with its neighbor. Solidity was expected to affect adaptation rate to subsequent script change. Within the Bayes' theorem framework, the goal probabilistic learning can be described as the acquisition of appropriate models for inference based on past experience. Events that cooccur persistently shape a solid model. The estimated likelihood of an event is dependent on its base-rate and how reliably it occurred in the past given that an associated event had happened. This likelihood is adapted on each iteration of the predictive and the associated event [Fiser et al., 2010]. The more often one event has followed another, the closer is the association between them and the more likely seems the succession. Hence, within solid models, the likelihood of the respective next event is very high. This tying of prediction to a conditional probability is proposed to result in slower adaptation of more solid models. It takes longer to rewrite, or rather rewire, strong associations. Lastly, we were interested in "biased" adaptation stages at early and advanced stages of learning. In biased stages, the number of iterations of divergent expositions differed considerably from the number of iterations of the respective original script. These states are of specific interest to the validation of predictions. To resurrect the picture outlined above, only a well-known path blocked/diverged instigates maintenance of the original idea, or "shielding" predictions from divergent influences. But previous experiences in a new environment should pale in insignificance to repeatedly coming across a divergence for the creation of an internal script and its predictions.

## Functional Neuroanatomy

As a main effect of the factor "adaptation," we expected adaptation of the internal model to the divergent script to lead to BOLD attenuation in a premotor-parietal network. **The premotor-parietal network** is associated with action observation and prediction of external events [cf., Schubotz, 2007]. **Its** parietal constituent is associated with coding for object pragmatics and space [Fagg and Arbib, 1998]. The frontal constituent, the lateral premotor cortex has been suggested to code for transformations underlying both our movements as well as observed events, for example changes in the position of objects [cf., Schubotz, 2007], and hence contributes to both action planning and action prediction. The concept of prediction refers to "filtering" of anticipated perception as has been described in motor control theories [Wolpert and Flanagan, 2001; cf., Schubotz, 2007]. We therefore expected that repeated exposure of the same action would lead to a decrease of activity in the premotor-parietal network, signifying adaptation.

As a main effect of the factor "solidity," we hypothesized higher activity for more solid compared to weaker models in the hippocampal formation. The close proximity of the concept of solidity to associative strength [Eichenbaum, 2000; Kim and Baxter, 2001; McClelland et al., 1995] and probabilistic learning [Kim and Baxter, 2001; Turke-Brown et al., 2010] points toward an involvement of the hippocampal cortex, revealed in stronger activity for more solid compared to less solid models [Eichenbaum, 2000; Kim and Baxter, 2001; McClelland et al., 1995; Turke-Brown et al., 2010].

Finally, we expected a significant interaction of the factors "solidity" and "adaptation." This common-sense assumption is supported by the fact that habits (also habits of thought), as an example for solid associations, are particularly difficult to unlearn [see Graybiel, 2008 for a review]. Moreover it has been established that stable environments, which by inference allow shaping solid models, are signified by a slow learning rate [Rushworth and Behrens, 2008]. However, as the neural correlates of an influence of solidity on adaptation have not been investigated so far, the study was explorative concerning the existence and location of the interaction's neural correlates.

## Implementation

To test our hypothesis, we familiarized participants previous to the fMRI session with a number of scripts containing everyday life actions, for example, a movie of making a salad. Each script encompassed a number of action steps for example, taking a bowl, grasping the lettuce, placing it in the bowl, sprinkling vinegar on top, taking salad tongues, tossing the salad. Original scripts were presented at three, six, or nine times in a preexperimental exposition session. In the fMRI session, participants encountered some scripts in the same version as before. Some scripts, however, the sequence changed from a certain point on. For example, the salad script now contained the subevents taking the bowl, grasping the lettuce, placing it in the bowl, reaching for the cheese, reaching for a knife, cutting pats of cheese into the bowl. Note that divergent scripts did not contain any action slips but were actions as valid as the original. Each script was shown nine times during the fMRI, either nine times in the original or nine times in the divergent version (no script appeared in two versions during the fMRI). Two main effects and their interaction were calculated:

1. To investigate the solidity effect, we contrasted the perception of divergent scripts with a large number (i.e., nine) of preexperimental expositions (factor level "solid") with the perception of divergent scripts with a low number (i.e., three) of preexperimental expositions (factor level "weak"). AQ3
2. To test whether adaptation would occur, we contrasted the first (i.e., first three—factor level "first") with the last (i.e., seventh to ninth) repetitions (factor level "last") of the divergent scripts pooled over all preexperimental exposition frequencies.

Finally, we aimed to establish a neuronal network that would reflect the dependence of adaptation rates on model solidity. To this end, we calculated the interaction contrast between the two-level factors "adaptation" and "solidity."

## METHODS

### Subjects

Nineteen right-handed, healthy participants (seven women, age 22–30 years old, mean age 25.3 years) took part in the study. The participants were right handed as assessed with the Edinburgh handedness inventory [Oldfield, 1971]. All participants were health screened by a physician and gave written informed consent.

### Stimuli and Task

The stimulus material contained 37 different movies of 8- to 12-s length. The movies were shot from the third-person perspective, not showing the actor's face. They contained every-day actions, taking place at a table. Most movie scripts, e.g., making a sandwich, existed in two versions (a and b). These scripts had an identical beginning, but started to diverge at some individual point, where after no commonality existed (Fig. 1). F1 Each version of each script was filmed 18 times. Thus, even though the same script appeared repeatedly during the preexperimental exposition and the experiment, the exact same shot of each script occurred only once. This method was employed to minimize surface-similarities between the scripts. A subset of 13 scripts was filmed in five different versions.

The experiment consisted of a preexperimental exposition of the movie material and an fMRI session starting exactly 15 min after the end of the preexposition. During the preexperimental exposition session, participants were seated in a sound attenuated chamber facing a computer screen. Distance to the screen was adjusted to ensure that the video displayed on the screen did not extend 5° of visual angle. They watched 27 scripts, a third of which was displayed three times, another third six times and the last third nine times in a randomized fashion over the course of 28-min lasting session. The participants saw one version of each script; but each repetition was another shot of the same script. Questions concerning whether some action or another was part of the script (e.g., "grasping an apple?") were posed on average after every fifth script to ensure ongoing attention to the stimulus material. Participants received visual feedback for 400 ms on whether they had answered correctly, incorrectly, or too late. After the preexposition, the participants were transferred directly to the fMRI chamber.

### FMRI Session

The fMRI session encompassed display of 36 different scripts. Each script was repeated nine times over the experiment. Nine scripts that had previously been displayed during the preexposition returned in the fMRI session in the same version as before ("originals" hereafter). Another nine of the preexperimentally shown scripts were presented in the fMRI session only in their complementary version ("divergents" hereafter) (Figs. 1 and 2). F2 The last nine scripts appeared in five different versions during the fMRI, each being displayed only once ("unpredictables" hereafter). The first third of the originals, the divergents and the unpredictables had previously been displayed three times each, the second third of all three kinds six times each, and the last third nine times each. Additionally, the design encompassed nine scripts that were completely new to the participants ("new originals") when they were displayed during the fMRI. The latter as well as the unpredictables will not be subject of the present paper but discussed in detail in a companion paper [Schiffer et al., in preparation]. However, the likely psychological effect of the unpredictables should be taken into account.

COLOR

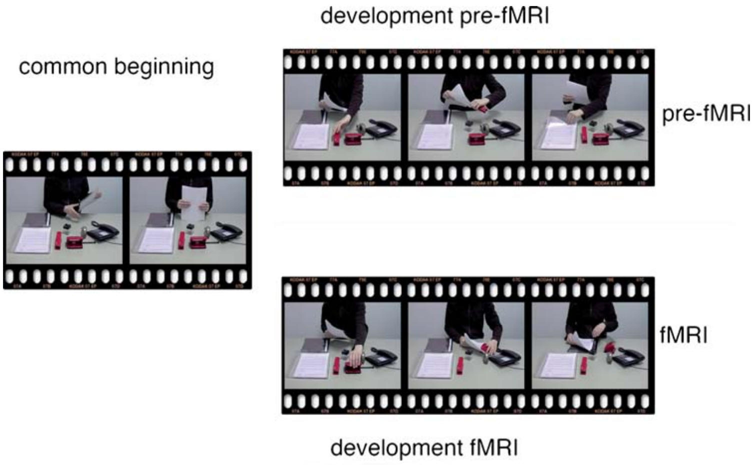


**Figure 1.**
The initial version that was displayed previous to the fMRI and the divergent version that was displayed during the fMRI had a common beginning, i.e., they started with the same action step(s).

Their presence and the associated experience of constantly changing scripts should decrease the likelihood of a divergent to be accepted as persistent at first encounter. That means that having seen a divergent only once does not allow the prediction that it returns in the same fashion—it could still turn out unpredictable at the second encounter. Only the second encounter of the same divergent delivers evidence that this script, albeit changed, is “learnable.”

The randomization distributed scripts of the same function, for instance the first presentation of the divergent version, evenly across the session. Thus, the temporal correlation between the function of a script and experiment duration, as well as the accumulation of identical functions during a specific period was minimized.

During the fMRI session, participants lay supine on the scanner bed. Their head and arms were stabilized using form-fitting cushioning and their hands rested on a rubber foam tablet. On the right hand side, a response panel was mounted on the tablet and fixed with tape. With their right hand index and middle finger resting on two response buttons, participants could answer the 32 intermittent questions concerning the content within the same F3 response-contingencies as in the preexposition (Fig. 3). Participants had three seconds to answer the question. Feedback on whether a response had been registered or not was displayed on the screen for 400 ms. The participants wore earplugs and headphones to attenuate scanner noise. Participants saw a reflection of the screen in a mirror, built into the head-coil and adjusted individually to allow for comfortable view of the entire screen. The movies did not extend further than 5° of visual angle in the mirror image of the computer screen. Sixteen null-events of 10-s length were displayed, consisting of display of the gray background on the screen. Participants were instructed to relax during null-events.

### Data Acquisition

The functional imaging session took place in a 3T Siemens Magnetom Trio scanner (Siemens, Erlangen, Germany). In a separate session, prior to the functional MRI, high-resolution 3D T-1 weighted whole-brain MDEFT sequences were recorded for every participant (128 slices, field of view 256 mm, 256 × 256 pixel matrix, thickness 1 mm, spacing 0.25 mm).

The functional session engaged a single-shot gradient echo-planar imaging (EPI) sequence sensitive to blood oxygen level-dependent contrast (28 slices, parallel to the bicommisural plane, echo time 30-ms, flip angle 90°; repetition time 2,000 ms; serial recording). Following the functional session immediately, a set of T1-weighted 2D-FLASH images was acquired for each participant (28 slices, field of view 200 mm, 128 × 128 pixel matrix, thickness 4 mm, spacing 0.6 mm, in-plane resolution 3 × 3 mm$^2$).

### FMRI Data Analysis

Functional data were offline motion-corrected using the Siemens motion protocol PACE (Siemens, Erlangen, Germany). Further processing was conducted with the LIPSIA software package [Lohmann, et al., 2001]. Cubic-spline interpolation was used to correct for the temporal offset

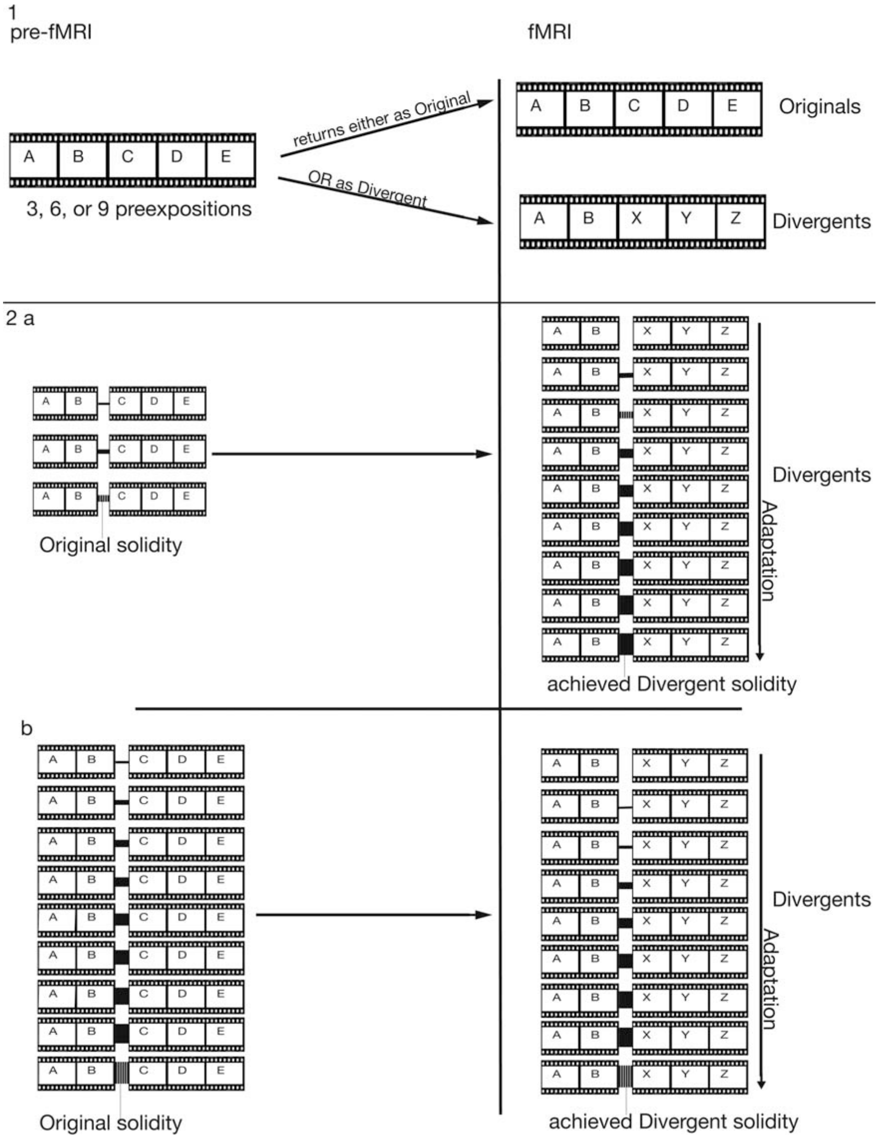
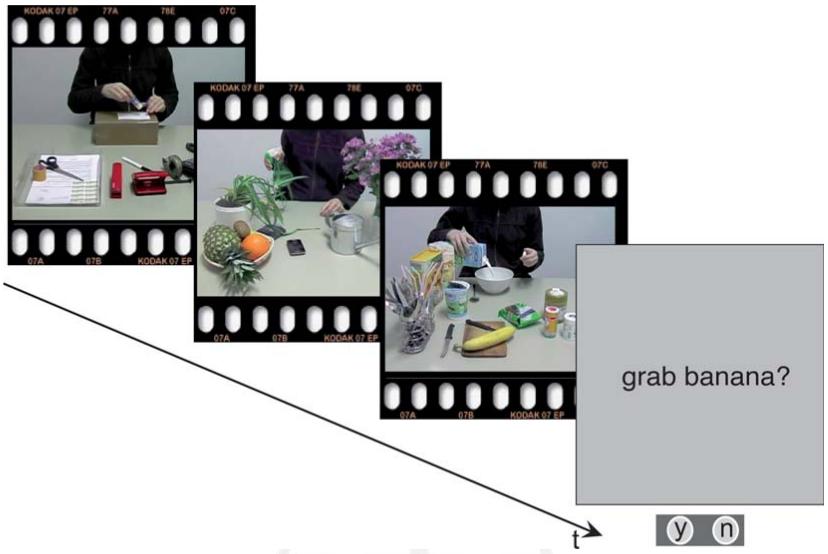


**Figure 2.**

Abstract representation of the script-structure. Letters refer to action steps. (1) Movies were preexposed three, six, or nine times in one version. A third of the movies reappeared in the fMRI in the same version as before "original." Another third appeared in a "divergent" version. This version started exactly as the original version had, but developed differently thereafter. (2a) Movies that were preexposed three times returned nine times as divergents during the fMRI. Strength of the indicated link reflects solidity; only the solidity of the transition of importance is indicated; each transition has the same assumed solidity in the beginning. (2b) Movies that were preexposed nine times similarly returned nine times as divergents during the fMRI. Again only the solidity of the relevant, i.e., later breached transition is graphically indicated.



COLOR

**Figure 3.**
During the fMRI session, participants watched divergents and originals in a randomized fashion and had to answer content-related questions on average after every 5th script.

between the slices acquired in one scan. To remove low-frequency signal changes and baseline drifts, a 1/110 Hz filter was applied. The matching parameters (six degrees of freedom, three rotational, three translational) of the T1-weighted 2D-FLASH data onto the individual 3D MDEFT reference set were used to calculate the transformation matrices for linear registration. These matrices were subsequently normalized to a standardized Talairach brain size [$x$ = 135 mm, $y$ = 175 mm, $z$ = 120 mm; Talairach and Tournoux, 1988] by linear scaling. The normalized transformation matrices were then applied to the functional slices, to transform them using trilinear interpolation and align them with the 3D reference set in the stereotactic coordinate system. The generated output had thus a spatial resolution of 3 × 3 × 3 mm$^3$.

The statistical evaluation was based on a least-square estimation using the general linear model (GLM) for serially autocorrelated observations [Worsley and Friston, 1995]. Temporal Gaussian smoothing (4 s FWHM) was applied to deal with temporal autocorrelation and determine the degrees of freedom [Worsley and Friston, 1995]. A spatial Gaussian filter of FWHM 5 mm was applied. The design matrix was generated by hemodynamic modeling using a ©-function and encompassed the first derivate. The onset vectors in the design matrix were modeled in a time-locked event-related fashion.

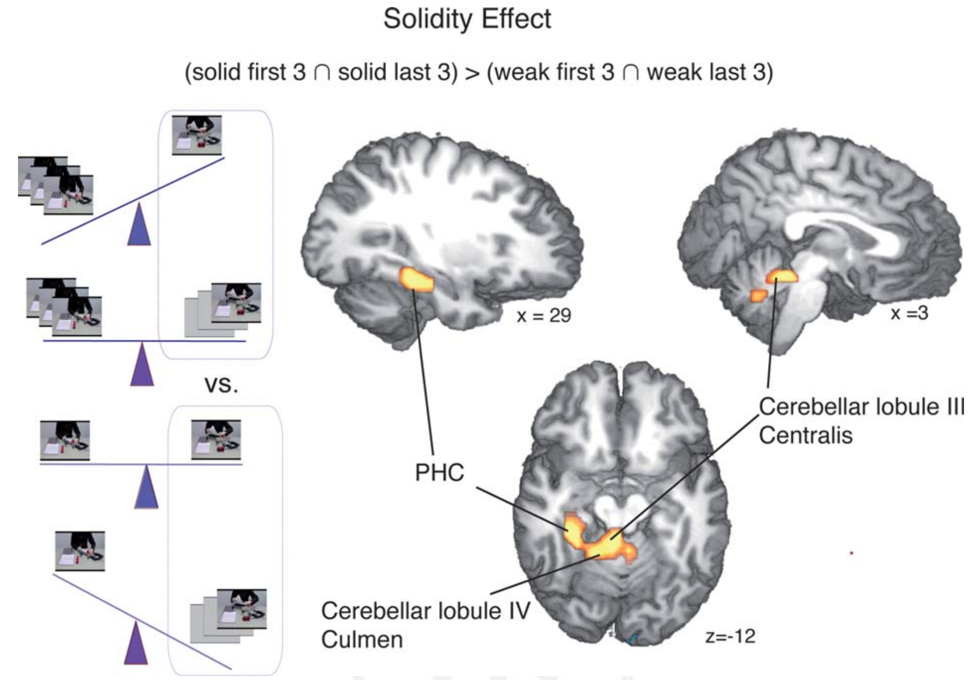
All contrasts were drawn from one design matrix. The first contrast accounted for the effect of model "solidity." The second contrast accounted for the overall adaptation effect. The third contrast targeted the interaction between model solidity and adaptation. To ensure that the activation from the interaction contrast was rooted in an orthogonal interaction, we also calculated the conjunction analysis that accounted for the same proposed interaction effect. The onset vectors were modeled to the point in time when the divergent was recognizable as divergent (hereupon "breach," Fig. 1). This breach had previously been visually timed to the moment when movement trajectories revealed that either the manipulation or the reached-for object was different from that in the originals. All divergents as well as the null-events were added as conditions of no-interest into the design matrix.

### Main effect Solidity

This effect was calculated as (solid / first ∩ solid / last) > (weak / last ∩ weak / first). Factor level "solid" refers to models that had been preexposed nine times; factor level "weak" refers to models that had been preexposed three times. Factor level "first" refers to first three presentations of a divergent; factor level "last" refers to its last three presentations (Fig. 4). F4

### Main effect adaptation

This effect was calculated as (solid / first ∩ weak / first) > (solid / last ∩ weak / last). Please refer above for explanation of the factor levels (Fig. 5). F5



COLOR

**Figure 4.**
The effect of model solidity was calculated contrasting the 1st to 3rd and 7th to 9th iteration of scripts that had been preexposed nine times with the 1st to 3rd and 7th to 9th iteration of scripts that had been preexposed three times. PHC: Parahippocampal cortex.

### *Interaction solidity by adaptation*

The interaction contrast signifies the interaction between the two two-level factors “solidity” and “adaptation,” and is thus derived from the crossing of the respective levels. Hence, it was calculated as contrast (solid / first > weak / first) > (solid / last > weak / last). Please refer above for F6 explanation of the factor levels (Fig. 6).

To enable an interpretation of the significant effects derived from this interaction contrast, it was important to ensure that all significant voxels reflected the same direction of the effect (this rationale applies to all interaction contrasts in fMRI). Therefore, we additionally calculated the conjunction of the contrasts (weak / first > weak / last) and (solid / first > solid / last).

All contrast images were fed into a second-level random effects analysis. The group analysis consisted of one-sample t tests across all contrast images to analyze whether the observed differences between conditions were significantly deviant from zero. Acquired t-values were transformed to $z$-scores. A two-step correction for false positive results based on a Monte-Carlo simulation was performed. In a first step, an initial $z$-threshold of 2.33 ($P < 0.05$, one-tailed) was applied to the simulated voxels. Afterward, based on the remaining clusters, statistically thresholds were calculated to correct for false positives at a significance level of $P = 0.05$. Cluster size as well as cluster value were taken into account at thresholding in a compensatory matter to prevent neglecting true positive activations in small anatomical structures [Lohmann et al., 2008]. Hence, all reported activations were significantly activated at $P \leq 0.05$, corrected for multiple comparisons at cluster level.

## Pilot Study

Previous behavioral results support the validity of the described contrasts. A preceding pilot study in another group of participants had provided behavioral evidence for the influence of solidity on adaptation. In the study, participants viewed each movie first three, six, or nine

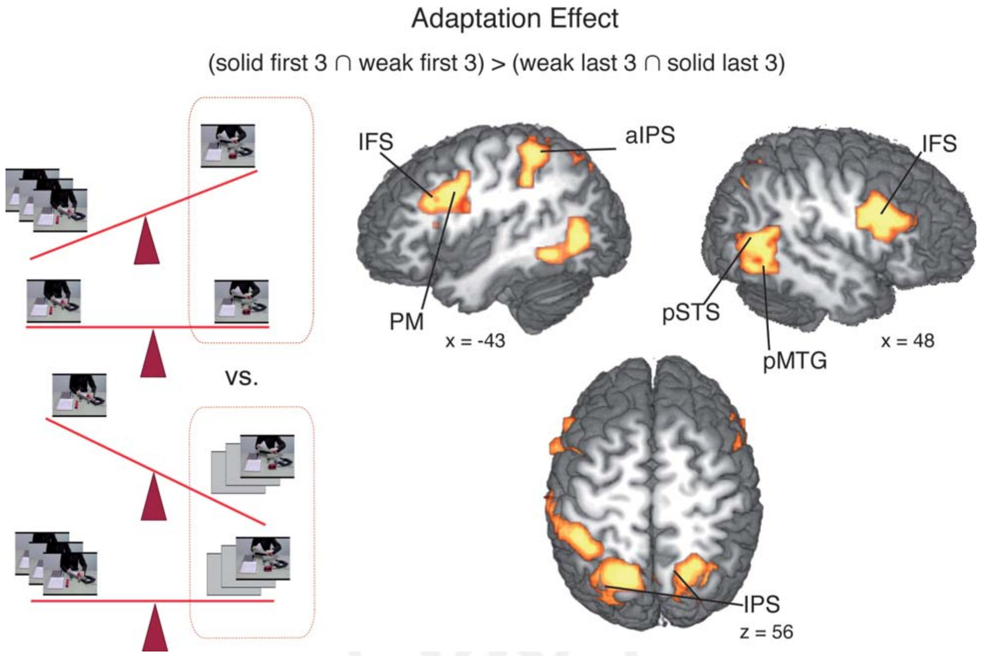


**Figure 5.**

The effect of model adaptation effect was calculated contrasting the 1st to 3rd iteration of scripts that had been preexposed either three or nine times with the 7th to 9th iteration of scripts that had been preexposed either three or nine times. (a) IPS: (anterior) intraparietal sulcus; IFS: inferior frontal sulcus; pMTG: posterior middle temporal gyrus; pSTS: posterior superior temporal sulcus; PM: premotor cortex.

times in the original version, followed by three, six, or nine divergent displays and eventually one or two original presentations. Meanwhile they had to constantly indicate whether the version that was on display at the moment was identical to the last display, or represented a change in script. We measured reaction times (RT) for the responses and conducted repeated measures ANOVA on the RTs of all correct responses to repetitions of divergents. The repeated measures ANOVA thus included two factors, the two-level factor original presentations (levels: three original presentations, nine original presentations) and eight-level factor divergent iteration (levels: 2nd ieration, 3rd iteration, . . . , 9th iteration). The first divergent was not included in the analysis, as it demanded a different response (indication of change) than the ensuing divergents (indication of repetition). The interaction effect between number of original presentations and iteration of the divergent approached significance at $P = 0.07$ (Greenhouse-Geisser corrected). To disentangle what effect carried the interaction we correlated the RT for each iteration with the number of previous originals. The correlation between RT of the divergents that had been displayed three times as original and their iterations was not significant ($r = 0.081$, $P = 0.3$). In contrast, the correlation between RT of the divergents that had been displayed nine times as original and their iterations approached significance ($r = -0.157$, $P = 0.06$). This marginal correlation reveals a continuous decrease in reaction times that we take to reflect ongoing adaptation to the divergents that had previously been shown nine times in their original version. Taken together, these results reflect a difference in adaptation rate dependent on the number of preexpositions.

## RESULTS

### Behavioral Results

The participants answered on average 87% of the 32 questions correctly (<27 questions). Standard deviation

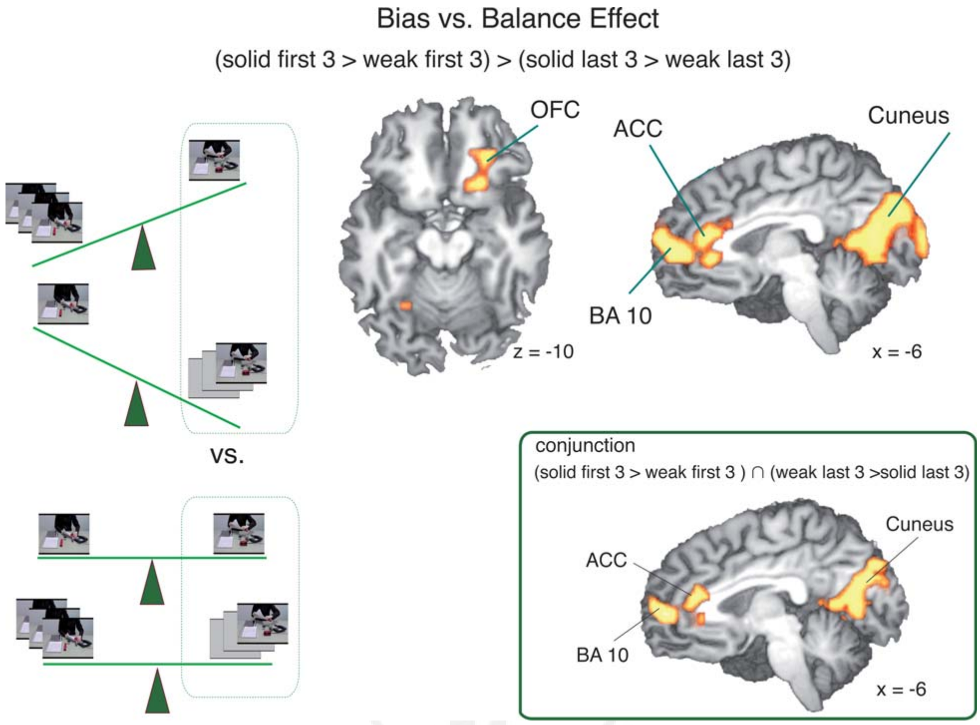

Bias vs. Balance Effect
(solid first 3 > weak first 3) > (solid last 3 > weak last 3)
OFC
ACC
Cuneus
BA 10
z = -10
x = -6
vs.
conjunction
(solid first 3 > weak first 3 ) ∩ (weak last 3 >solid last 3)
Cuneus
ACC
BA 10
x = -6

**Figure 6.**

The biased vs. balanced effect was calculated contrasting the 1st to 3rd iteration of scripts that had been preexposed nine times and the 7th to 9th iteration of scripts that had been preexposed three times with the 1st to 3rd iteration of scripts that had been preexposed three times and 7th to 9th iteration of scripts that had been preexposed nine times. ACC: anterior cingulate cortex; BA 10: Brodmann area 10; OFC: orbitofrontal cortex.

was 7%. In the postexperimental questionnaire participants were asked whether all movies had returned as before and no participant indicated that all movies had. Six of the 19 participants reported spontaneously to the open question whether they wished to report anything whatsoever, that some movies were different than before. This behavioral measure furthers the argument that the participants were aware that some movies were altered versions of what they had seen preexperimentally, instead of believing that the different movies (divergents) were not related to the initial version.

## FMRI Results

The model "solidity contrast" (solid / first ∩ solid / last) > (weak / last ∩ weak / first) yielded activity in the right parahippocampal cortex, and also in the right cerebellar Lobule III (centralis) and bilaterally in the Lobule IV (culmen) of the cerebellum (Table I) (Fig. 4).

T1

The model "adaptation contrast" (solid / first ∩ weak / first) > (solid / last ∩ weak / last) yielded bilateral

**TABLE I. Solidity contrast: Anatomical specification, Talairach coordinates (*x,y,z*) and maximal *Z*-scores of significantly activated voxels for model solidity: divergents with high (nine preexpositions) or weak (three preexpositions) model solidity**

| | Talairach coordinates | | | Z-values, local |
|---|---|---|---|---|
| Localization | *x* | *y* | *z* | maxima |
| Parahippocampal cortex | 32 | −32 | −12 | 3.43 |
| Cerebellum, Lobule III, Centralis | 4 | −38 | −9 | 5.16 |
| Cerebellum, Lobule IV, Culmen | −8 | −47 | −18 | 4.8 |

**TABLE II. Adaptation contrast: Anatomical specification, Talairach coordinates (*x*,*y*,*z*) and maximal *Z*-scores of significantly activated voxels for the model adaptation: first vs. last presentations of divergents**

| | Talairach coordinates | | | Z-values, local |
|---|---|---|---|---|
| Localization | *X* | *y* | *z* | maxima |
| Superior parietal lobule | −14 | −59 | −57 | 3.67 |
| Intraparietal sulcus | 32 | −62 | 45 | 4.18 |
| | −20 | −65 | 39 | 3.74 |
| | −40 | −41 | 54 | 3.6 |
| Intraparietal sulcus, anterior segment | −58 | −23 | 42 | 2.9 |
| Premotor cortex | −46 | 10 | 24 | 3.69 |
| Inferior frontal gyrus | 46 | 16 | 30 | 3.72 |
| | −44 | 22 | 24 | 3.16 |
| Posterior middle temporal gyrus | 44 | −56 | 15 | 3.72 |
| | 40 | −47 | −3 | 3.81 |
| | −46 | −65 | 12 | 3.25 |
| | −40 | −50 | −6 | 3.4 |

activity in the inferior frontal sulcus (IFS), the left premotor cortex (PM), the left superior parietal lobe (SPL), and intraparietal sulcus (IPS), extending into anterior IPS in the left hemisphere. The posterior middle temporal gyrus T2 (MTG) was activated bilaterally (Table II) (Fig. 5).

The solidity by adaptation interaction contrast (solid / first > weak / first) > (solid / last > weak / last) showed significant activation of the frontopolar cortex comprising mesial Brodmann Area (BA) 10 and right lateral BA10. Further activations were in the anterior cingulate cortex (ACC), right orbitofrontal cortex (OFC), the right striatum, right posterior superior temporal gyrus (pSTS), cuneus, T3 and the left fusiform gyrus (Table III; Fig. 6). The second approach to this analysis, the conjunction analysis (iii-a), i.e., (weak / last > weak / first) ∩ (solid / first > solid / last), yielded activity in the mesial and the lateral BA10, ACC, and cuneus, and in the right fusiform gyrus T4 (Table IV).

## DISCUSSION

Internal models of an action encompass expectations on the development of this action [Bar, 2009; Jeannerod, 1995]. Valid predictions make perception more efficient and are beneficial to fast reactions [Wolpert and Flanagan, 2001]. The present fMRI study investigated the neural correlates of the influence of the solidity of the original internal model of an action on subsequent adaptation of the internal model to a divergent script. To that end, participants watched movies that familiarized them with the original scripts and thus to establish internal model of them. In the fMRI they were confronted with divergent versions of the previously learnt scripts.

We found a persistent effect of preexperimental exposition frequency (main effect of solidity) in the right parahippocampal cortex as implied by the concept's proximity to associative strength. There was also an effect of solidity bilaterally in the anterior cerebellum. This result stresses the importance of previous experience to expectation, especially in the face of new information. As hypothesized, divergent experiences incited adaptation in fronto-parietal motor regions, i.e., left PMv, bilateral IFS and IPS. Moreover the adaptation effect was evident in the posterior MTG and in the left SPL. Finally, the exciting finding of a network dealing with a solidity bias, i.e., stages where solidity of one script surpasses that of another (solidity by adaptation interaction), supports the notion of a lasting influence of possible alternatives. The activity that was found for this interaction, located in the left frontomedian cortex (FMC), i.e., BA 10 and the ACC, as well as right striatum and right OFC, suggests a continuous processing of divergent information in these areas, be it current or past.

### Solidity Exerts Prolonged Influence

Activity in the solidity contrast reflects an ongoing response to divergent scripts that is more pronounced for solid than for weaker original internal models. The cerebellar activity was in a classical motor region [Marvel and Desmond, 2010], in Lobules III and IV [Schmahman et al., 1999]. Working memory function, proposed for cerebellar Lobules VI/Crus I [Marvel and Desmond, 2010] is rather an unlikely explanation for this anterior activity. Hence, we take it to reflect continuing mismatch between the internal motor model's expectations and perception, which is increased if the original internal model was highly habituated. The parahippocampal cortex has been associated with topographical learning [Aguirre et al., 1996], scene processing [Epstein and Kanwisher, 1998] and the

**TABLE III. Interaction contrast: Anatomical specification, Talairach coordinates (*x*,*y*,*z*) and maximal *Z*-scores of significantly activated voxels for biased vs. balanced states: the first divergents of a solid internal model and the last divergents of a weak internal model vs. the first divergents of a weak internal model and the last divergents of a solid internal model**

| | Talairach coordinates | | | Z-values, local |
|---|---|---|---|---|
| Localization | *X* | *y* | *z* | maxima |
| Frontal pole, BA10 | −10 | 61 | 12 | 4.31 |
| | 14 | 52 | 9 | 3.33 |
| Anterior cingulate gyrus, BA24 | 2 | 34 | 15 | 2.85 |
| | −4 | 31 | 15 | 2.79 |
| Orbitofrontal gyrus | 22 | 31 | −9 | 3.14 |
| Cuneus | 8 | −77 | 18 | 3.81 |
| Posterior superior temporal sulcus | 56 | −32 | 9 | 3.8 |
| Fusiform gyrus | −26 | −56 | −6 | 3.19 |
| Striatum | 20 | 19 | −3 | 4.1 |

**TABLE IV. Conjunction analysis: Anatomical specification, Talairach coordinates (*x*,*y*,*z*) and maximal *Z*-scores of significantly activated voxels for biased vs. balanced states: the first divergents of a solid internal model vs. the first divergents of a weak internal model and the last divergents of a weak internal model vs. and the last divergents a solid internal model**

| | Talairach coordinates | | | Z-values, |
|---|---|---|---|---|
| Localization | *X* | *y* | *z* | local maxima |
| Frontal pole, BA10 | 6 | 43 | 3 | 2.40 |
| | −4 | 49 | 3 | 2.91 |
| Anterior cingulate gyrus, BA24 | 2 | 31 | 15 | 3.54 |
| | 2 | 34 | −3 | 2.16 |
| | −4 | 28 | 0 | 3.89 |
| Cuneus | −2 | −71 | 21 | 2.83 |
| Fusiform gyrus | 16 | −53 | −6 | 2.46 |

association of scenes and locations with objects [Bar et al., 2008; Sommer et al., 2005]. Here, we propose that parahippocampal activity signifies the revision of associations [Eichenbaum, 2000; McClelland et al., 1995] between scenes and actions or action-relevant objects. The present data allow no decision between these alternatives as the divergent script sometimes included the use of a different object than the original script did, but sometimes only entailed an altered manipulation of the same object.

### Adaptation in the Cortical Motor Network

The adaptation contrast (ii) yielded activity in the left PM(v), the bilateral IPS and the left posterior MTG, a network that is not only relevant for action execution, but also prominent in action observation [Jeannerod, 1995]. The adaptation contrast tested whether the hypothesized fronto-parietal motor regions would be sensitive to violated expectations and show an adaptation to the new action script.

During the first encounters of the divergent script, perception was assumed to deviate from the internal model. An increase of neuronal activity at this stage reflects a breach of expectation signal that incites learning [Summerfield et al., 2008]. This signal can also be understood as a correlate of the processing of unexpected (salient) objects or manipulations [Keysers and Perret, 2004]. These functions can be seen as two sides of the same coin. Accordingly, the original script acts like a filter that minimizes processing demands of all according perceptions. Divergent perceptions, however, are not filtered, rendering them more salient than prefiltered perceptions. The resulting increased activation is a "breach of expectation signal" and incites learning. As soon as the divergent script has been learnt, it can serve as a filter for all according perceptions again.

Adapting the internal model to account for the divergent script is a learning or relearning process, and in a stable environment, strong evidence should be required to motivate learning [Rushworth and Behrens, 2008]. Otherwise, assembling and memorizing experiences would be pointless, as they would loose their capacity to guide successful behavior as soon as a one-time breach of expectation had occurred. Hence, the divergent perception should not cause instantaneous adaptation of the internal model; accordingly, a process of adaptation is revealed by diminution of the neural correlate of divergence over a large number of iterations [Friston et al., 2006; Grill-Spector et al., 2006; Majdanžić et al., 2009] as targeted in the adaptation contrast (ii). It has previously been established that the cortical motor network is capable of predicting the ongoing course of action [Jeannerod, 1995]. The current study furthers our understanding thereof, suggesting that the network is sensitive to salient violations of its predictions and shows appropriately slow adaptation. A detailed account of the proposed functions of the constituents adapting in this process will be supplied below.

The SPL has been discussed as a potential site of spatial priority maps, which designate relevant object locations and can be internally guided or externally cued [Molenberghs et al., 2007; Nobre et al., 2004]; one of the SPL's functions seems to be constructing and changing these spatial priority maps [Chiu and Yantis, 2009; Molenberghs et al., 2007]. Activity in the adaptation contrast is evidence for the remapping of spatially guided attention in SPL; this remapping or changing of weights in the priority map [Molenberghs et al., 2007] becomes important to action emulation as suddenly relevant objects demand attention, while previously used objects loose their significance for the action sequence.

Activity in the posterior MTG is taken to reflect increased processing of the movements of the actor and the actions associated with suddenly relevant objects [Beauchamp and Martin, 2007; Beauchamp et al., 2002]. Divergent scripts encompassed use (and accordingly motion) of different objects or different use of the same object as the original scripts. Encounter of the first presentations of the divergent script entailed a mismatch between emulated associations and valid, but unpredicted perceived use. Activity in the posterior MTG has been discussed in association with the frontoparietal motor network [Beauchamp and Martin, 2007; Johnson-Frey, 2004]. The role of this frontoparietal network of IPS and PM in goal-directed object manipulation and internal modeling thereof has been researched extensively [Grèzes and Decety, 2001; Jeannerod, 1995; Johnson-Frey, 2004 for reviews]. The anterior IPS has been proposed to provide the ventral premotor cortex with information on object pragmatics [Fagg and Arbib, 1998; Schubotz and von Cramon, 2008]. Attenuation of its activity has previously been interpreted as a teaching signal that allows model adaptation [Tunik et al., 2007]. Medial IPS has previously been reported to be crucial for the online control of goal-directed precision movement [Grefkes and Fink, 2005 for a review]. Online correction relies on the detection of mismatch between internal emulation and sensorimotor information [Wolpert and Flanagan, 2001]. We suggest that the activity along IPS reflects a decreasing mismatch between the internal model's emulated action and the currently perceived action. The closely linked [Geyer et al., 2000] PMv, which is assumed to store action knowledge and object function, shows increased activity when new scripts have to be learnt [see Schubotz and von Cramon, 2003 for review]. Activity in premotor cortex is increased when prediction [Schubotz and von Cramon, 2003], or simulation [Grèzes and Decety, 2001], and planning of movements [Johnson-Frey, 2004] is involved. Against this backdrop, PMv activation during the first encounters of unpredicted divergences can be regarded as further evidence of this area's involvement in compiling complex actions.

*Initial bias toward the original script*

Activity in IFS has been suggested to modulate the bias between competing representations [Badre et al., 2005; AQ4 Kuhl et al., 2007; Wurm and Schubotz, unpublished data]. This fits well with an influential model of prefrontal cortex function that suggests that prefrontal cortex is involved in activating and supporting relevant but unfavored or weak associations [Miller and Cohen, 2001**]**. The present study delivers new evidence for the assumption that the IFS support weak models: attenuation of IFS activity points to its involvement in supporting the new divergent internal model and its associations during the first encounters of the divergent script. Each iterations of this divergent script should solidify its representation, diminishing IFS activity as a balanced state of competition between original and divergent internal model is approached and the bias runs eventually in favor of the new internal model [Schubotz and von Cramon, 2008].

### Bias vs. Balance—Prefrontally Mediated Integration of Incompatible Models

The activation of the FMC, occipital areas, as well as the pSTS in the solidity–adaptation interaction contrast revealed these areas' involvement in processing information when the solidity of one internal model surpasses that of another. Strikingly, this network was found to be involved not only when this bias run in favor of the original script (and hence, against the currently perceived one), but also when the bias was already toward the actually presented action (and hence, against the former original script). The underlying analysis was explorative concerning the areas that would be involved in the interaction of solidity and adaptation. However, the interesting results help to explain previous puzzling findings [Frank et al., 2005] and enhance our understanding of a conundrum in the EEG-centered conflict-monitoring literature:

FMC activity spread from the ACC into BA10. The ACC is understood to be responsive to bias, especially in decision and stimulus–response paradigms [Bunge et al., 2004; Miller and Cohen, 2001]. It is supposed to convey this bias to the dorsolateral prefrontal cortex [Miller and Cohen, 2001]. Classic bias-related responses recorded in the ACC focus on conflict [see Botvinick et al., 2004; van Veen and Carter, 2002 for review]. Conflict is often understood as bias running against the necessary association, demanding PFC to support or maintain activation of a "weaker" association [Kuhl et al., 2007; Miller and Cohen, 2001]. This "conflict solving," triggered by the ACC, could also mean suppression of an unlikely target [Kuhl et al., 2007], apart from the classic conception as fostering a weaker alternative [Miller and Cohen, 2001]. The current study, in contrast, revealed that the ACC is active for both biased states, even when perception is in accordance with the currently more solid internal representation. This latter form of bias, however, is not signified by what is often understood as conflict, i.e., the need to resolve competition in favor of the weaker alternative. Consequently, IFS activation is diminished at this stage, as apparent in the adaptation contrast and discussed above, while it is present when bias does run against the presented model at the beginning of adaptation. The proposed bias account is in line with an account of ACC function that integrates conflict monitoring and more general evaluative computation [Botvinick et al., 2004]. Conflict would then mean the activation of the representations of two incompatible (action) models [Botvinick et al., 2004]. The present results seem to singularly underpin a point in the EEG literature of conflict monitoring with fMRI-derived results. Yeung et al. [2004] argue that the N2 component in correct trials and ERN component following errors is elicited when evidence for one representation outweighs that for the other—with the N2 preceding correct responses and the ERN being a posterror correlate of surmounting evidence for the (discarded) correct response. This aspect of "outweighing" the competing alternative, or bias, has however not always been taken into consideration in the conflict monitoring literature even though one study [Frank et al., 2005] found that in a forced choice task, a higher discrepancy between the respective reward values of two options resulted in a higher ERN than a more equal distribution of reward. Our study reveals that activity in the FMC is stronger if evidence is biased in favor of one of the incompatible representations, indicating in this case a higher predictive capacity for one model than the other. The study thus contributes to the clarification of the EEG centered conflict monitoring debate [Botvinick et al., 2004], corroborating a bias-related definition of conflict, as opposed to the notion of equally strong competitors.

The ACC is closely linked to BA10 [Allman et al., 2002]. A special kind of neuron, the spindle neurons in the ACC have been proposed to convey the motivation to adapt to changes to BA10 [Allman et al., 2002]. More generally, the frontopolar area is part of the hippocampal-cortical memory system [Vincent et al., 2008]. Moreover, BA10 is taken to be responsible for the integration of separate cognitive operations [see Ramnani and Owen, 2004 for review]. One example is episodic retrieval and success monitoring, a process that can be understood in terms of comparing an internal representation to an outcome [Ramnani and Owen, 2004]. We propose that only the biased states entailed suppression of either the original or the divergent internal model, respectively. The deterministic nature of the paradigm suggested solidifying the divergent internal model, thus the biased and balanced states both encompassed a need to register and to encode the divergent internal model. But the biased states also suggested suppression of either the original or the divergent. If there were no suppression of the divergent internal model in the beginning, learning would be instantaneous. This was not the case. If the diversion was not registered, accumulating evidence would not be tracked and learning would never set in. Once evidence for the validity of the divergent internal model outweighs that for the original, suppression of the neglected alternative is regarded as efficient [Kuhl et al., 2007] and guides expectations toward the most likely outcome. A coupling of the ACC and BA10 during suppression has previously been reported by Kuhl et al. [2007]. In the balanced states, evidence for neither internal model outweighs evidence for the other and suppression could be regarded as too persistent (for the divergent internal model) or too premature (for the original internal model), respectively.

Activity of the OFC in the interaction contrast complements the emerging picture [Ghahremani et al., 2009]. Biased states necessarily have one strong, or solid component, like a prepotent response or well practiced forward model. As discussed above, this strong component can trigger suppression of alternatives as it allows generation of hypotheses. Both, hypothesis generation and suppression have been discussed as potential OFC functions [Elliott et al., 2000; Ghahremani et al., 2009; Vartanian and Goel, 2005]. Hypothesis generation and suppression can be reframed as evaluation or weight changes, as a result of evaluation, which itself is a function ascribed to the OFC [Wallis, 2007]. A steady environment, as signified by the existence of one solid internal model, makes it worthwhile to track contingencies and integrate outcome histories into learning [Rushworth and Behrens, 2008]. Responses to contingency differences, another type of evaluation, have similarly been allocated in the OFC [Windmann et al., 2006]. We propose that the activity increase in the OFC during a state of bias is indicative of the evaluation of the current forward model [Schubotz and von Cramon, 2008] against the backdrop of one solid and one weak or paling internal model. Closely linked to the OFC in its evaluative function is the striatum that was similarly active in the interaction contrast [Grinband et al., 2006; Oenguer et al., 2003; Schoenbaum et al., 2009].

To sum up, the similarities the networks display during the beginning and during an advanced state of learning single model solidity bias out as the determinant factor, as opposed to conflict between equally strong representations. It is likely that there is only consolidation in the balanced state, but an integration of consolidation of one and suppression of the other internal model in the biased states. Thus, bias incites the same operation in different situations, i.e., suppression of the divergent internal model in the beginning and suppression of the original internal model in the end. In the beginning, the divergent script stands in stark contrast to a solid internal model with identical onset phases; hence, it demands attention [Summerfield, 2008], possibly against a backdrop of previous suppression. In the end, even though the old original internal model has not been valid for a large number of iterations, it still exerts an influence on predictions. The emergence of significant bias-related activations suggests that the opposite, i.e., a state of balance or ambiguity, is reached when the number of expositions of the divergent script matches the number of previous expositions of the original script. This finding is indicative of a slower adaptation rate for a solid, compared to a weak internal model and supported by the data from the pilot study (see Pilot Study section).

## Concluding Remarks

In a dynamic environment, it is particularly important not only to set up internal models but also to keep them up to date. Hence, expectations must be revised if they do not accord to our last experiences. However, unwarranted revision should be prevented, to not loose the gain of experience. The current study provided evidence for the notion that familiarity with an event influences the adaptation rate of according expectations.

## REFERENCES

Aguirre GK, Detre JA, Alsop DC, D'Esposito M (1996): The parahippocampus subserves topographical learning in man. Cereb Cortex 6:823–829.

Allman J, Hakeem A, Watson K (2002): Book review: Two phylogenetic specializations in the human brain. Neuroscientist 8:335–346.

Badre D, Poldrack RA, Parè-Blagoev EJ, Insler RZ, Wagner AD (2005): Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. Neuron 47:907–918.

Bar M (2009): The proactive brain: Memory for predictions. Philos Trans R Soc B Biol Sci 364:1235–1243.

Bar M, Aminoff E, Schacter DL (2008): Scenes unseen: The parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se. J Neurosci 28:8539–8544.

Beauchamp MS, Martin A (2007): Grounding object concepts in perception and action: Evidence from FMRI studies of tools. Cortex (a journal devoted to the study of the nervous system and behavior) 43:461–468.

Beauchamp MS, Lee KE, Haxby JV, Martin A (2002): Parallel visual motion processing streams for manipulable objects and human movements. Neuron 34:149–159.

Botvinick MM, Cohen JD, Carter CS (2004): Conflict monitoring and anterior cingulate cortex: An update. TrendsCogn Sci 8:539–546.

Bunge SA, Burrows B, Wagner AD (2004): Prefrontal and hippocampal contributions to visual associative recognition: Interactions between cognitive control and episodic retrieval. BrainCogn 56:141–152.

Caspers S, Zilles K, Laird AR, Eickhoff SB (2010): ALE meta-analysis of action observation and imitation in the human brain. NeuroImage 50:1148–1167.

Chiu Y-C, Yantis S (2009): A domain-independent source of cognitive control for task sets: Shifting spatial attention and switching categorization rules. J Neurosci 29:3930–3938.

Eichenbaum H (2000): A cortical-hippocampal system for declarative memory. Nat Rev Neurosci 1:41–50.

Elliott R, Dolan RJ, Frith CD (2000): Dissociable functions in the medial and lateral orbitofrontal cortex: Evidence from human neuroimaging studies. Cereb Cortex 10:308–317.

Epstein R, Kanwisher N (1998): A cortical representation of the local visual environment. Nature 392:598–601.

Fagg AH, Arbib MA (1998): Modeling parietal-premotor interactions in primate control of grasping. Neural Netw 11:1277–1303.

Fiser J, Berkes P, Orbán G, Lengyel M (2010): Statistically optimal perception and learning: From behavior to neural representations. Trends Cogn Sci 14:119–130.

Frank MJ, Woroch BS, Curran T (2005): Error-related negativity predicts reinforcement learning and conflict bias. Neuron 47:495–451.

Friston K, Kilner J, Harrison L (2006): A free energy principle for the brain. J Physiol Paris 100:70–87.

Geyer S, Matelli M, Luppino G, Zilles K (2000): Functional neuroanatomy of the primate isocortical motor system. Anat Embryol 202:443–474.

Ghahremani DG, Monterosso J, Jentsch JD, Bilder RM, Poldrack RA (2009): Neural components underlying behavioral flexibility in human reversal learning. Cereb Cortex. AQ5

Graybiel AM (2008): The basal ganglia: Learning new tricks and loving it. Curr Opin Neurobiol 15:638–644.

Grèzes J, Decety J (2001): Functional anatomy of execution, mental simulation, observation, and verb generation of actions: A meta-analysis. Hum Brain Mapp 12:1–19.

Grill-Spector K, Henson R, Martin A (2006): Repetition and the brain: Neural models of stimulus-specific effects. Trends Cogn Sci 10:14–23.

Grinband J, Hirsch J, Ferrera VP (2006): A neural representation of categorization uncertainty in the human brain. Neuron 49:757–763.

Jeannerod M (1995): Mental imagery in the motor context. Neuropsychologia 33:1419–1432.

Johnson-Frey SH (2004): The neural bases of complex tool use in humans. Trends Cogn Sci 8:71–78.

Keysers C, Perrett DI (2004): Demystifying social cognition: A Hebbian perspective. Trends Cogn Sci 11:501–507.

Kilner JM, Friston KJ, Frith CD (2007): Predictive coding: An account of the mirror neuron system. Cogn Process 8:159–166.

Kim JJ, Baxter MG (2001): Multiple brain-memory systems: The whole does not equal the sum of its parts. Trends Neurosci 24:324–330.

Kuhl BA, Dudukovic NM, Kahn I, Wagner AD (2007): Decreased demands on cognitive control reveal the neural processing benefits of forgetting. Nat Neurosci 10:908–914.

Lohmann G, Mueller K, Bosch V, Mentzel H, Hessler S, Chen L, et al. (2001): Lipsia—A new software system for the evaluation of functional magnetic resonance images of the human brain. Comput Med Imaging Graph Off J Comput Med Imaging Soc 25:449–457. AQ7

Majdanžić J, Bekkering H, van Schie HT, Toni I (2009): Movement-specific repetition suppression in ventral and dorsal premotor cortex during action observation. Cereb Cortex 19:2736–2745.

Marvel C, Desmond J (2010): Functional topography of the cerebellum in verbal working memory. Neuropsychol Rev 20:271–279.

McClelland JL, McNaughton BL, O'Reilly RC (1995): Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychol Rev 102:419–457.

Miller EK, Cohen JD (2001): An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Molenberghs P, Mesulam MM, Peeters R, Vandenberghe RRC (2007): Remapping attentional priorities: Differential contribution of superior parietal lobule and intraparietal sulcus. Cereb Cortex 17:2703–2712.

Neal A, Kilner JM (2010): What is simulated in the action observation network when we observe actions? Eur J Neurosci 32:1765–1770.

Nobre AC, Coull JT, Maquet P, Frith CD, Vandenberghe R, Mesulam MM (2004): Orienting attention to locations in perceptual versus mental representations. J Cogn Neurosci 16:363–373.

Oenguer D, Ferry AT, Price JL (2003): Architectonic Subdivision of the Human Orbital and Medial Prefrontal Cortex, Vol. 460. New York, NY: ETATS-UNIS, Wiley-Liss. AQ8

Ramnani N, Owen AM (2004): Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. Nat Rev Neurosci 5:184–194.

Rushworth MFS, Behrens TEJ (2008): Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci 11:389–397.

Schoenbaum G, Roesch MR, Stalnaker TA, Takahashi YK (2009): A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. Nat Rev Neurosci 10:885–892.

Schubotz RI (2007): Prediction of external events with our motor system: Towards a new framework. Trends Cogn Sci 11:211–218.

Schubotz RI, von Cramon DY (2003): Functional-anatomical concepts of human premotor cortex: Evidence from fMRI and PET studies. NeuroImage 20 (Suppl 1):S120–S131.

Schubotz RI, von Cramon DY (2008): The case of pretense: Observing actions and inferring goals. J Cogn Neurosci 21:642–653.

Schutz-Bosbach S, Prinz W (2007): Prospective coding in event representation. Cogn Process 8:93–102.

Sommer T, Rose M, Gläscher J, Wolbers T, Büchel C (2005): Dissociable contributions within the medial temporal lobe to encoding of object-location associations. Learn Mem 12:343–351.

Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T (2008): Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci 11:1004–1006.

Tunik E, Rice NJ, Hamilton A, Grafton ST (2007): Beyond grasping: Representation of action in human anterior intraparietal sulcus. NeuroImage 36 (Suppl 2):T77–T86.

Turke-Brown NB, Scholl BJ, Johnson MK, Chun M (2010): Implicit perceptual anticipation triggered by statistical learning. J Neurosci 30:11177.

van Veen V, Carter CS (2002): The anterior cingulate as a conflict monitor: fMRI and ERP studies. Physiol Behav 77:477–482.

Vartanian O, Goel V (2005): Task constraints modulate activation in right ventral lateral prefrontal cortex. NeuroImage 27:927–933.

Vincent JL, Kahn I, Snyder AZ, Raichle ME, Buckner RL (2008): Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. J Neurophysiol 100:3328–3342.

Windmann S, Kirsch P, Mier D, Stark R, Walter B, Guentuerkuen O, et al. (2006). On framing effects in decision making: Linking lateral versus medial orbitofrontal cortex activation to choice outcome processing. J Cogn Neurosci 18:1198–1211.

Wolpert DM, Flanagan JR (2001). Motor prediction. Curr Biol 11:729–732.

Worsley KJ, Friston KJ (1995): Analysis of FMRI time series revisited—again. Neuroimage 2:173–181.

Zacks JM, Speer NK, Swallow KM, Braver TS, Reynolds JR (2007): Event perception: A mind-brain perspective. Psychol Bull 133:273–293.

## 2.3 Surprised at all the Entropy: Hippocampal, Caudate and Midbrain Contributions to Learning from Prediction Errors

# PLoS ONE

## Surprised at all the Entropy: Hippocampal, Caudate and Midbrain Contributions to Learning from Prediction Errors

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | |
| **Article Type:** | Research Article |
| **Full Title:** | Surprised at all the Entropy: Hippocampal, Caudate and Midbrain Contributions to Learning from Prediction Errors |
| **Short Title:** | Prediction Errors in Striatum and Hippocampus |
| **Corresponding Author:** | Anne-Marike Schiffer<br>Max Planck Institute for Neurological Research<br>Cologne, NRW GERMANY |
| **Keywords:** | fMRI; hippocampus; caudate nucleus; Basal Ganglia; dopamine; action observation; associative mismatch; predictive coding; prediction error; Shannon entropy; surprise; action-observation |
| **Abstract:** | Influential concepts in neuroscientific research cast the brain a predictive mechine that revises its predictions when they are violated by sensory input. While this is famously implemented in the predictive coding account of perception, it also relates to learning. Learning from prediction errors however has been suggested for the hippocampal memory system as well as for the basal ganglia. The present fMRI study used an action-observation paradigm to investigate the contributions of the hippocampus, caudate nucleus and midbrain dopaminergic system to different types of learning: learning in the absence of predicton errors, learning from prediction errors, and responding to the accumulation of prediction errors in unpredictable stimulus configurations. We conducted analyses of the regions of interests' BOLD response towards these different types of learning, implementing a bootstrapping procedure to correct for false positives. We found both, caudate nucleus and the hippocampus to be activated by perceptual prediction errors. The hippocampal responses seemed to relate to the associative mismatch between a stored representation and new sensory input. Moreover, its response was significantly influenced by the average information, or entropy of the stimulus material. Behavioural measures indicated, that memory for information received under increasing entropy is worse than memory for stable representations. The habenula mirrored the caudate's responses to perceptual prediction errors unrelated to reward. Lastly, we found that the substantia nigra diplays a disparate response pattern: it was activated by the novelty of sensory input.<br>In sum, we established differential involvement of the hippocampus, caudate nucleus and midbrain dopaminergic system in different types of learning. We relate learning from perceptual prediction errors to the concept of predictive coding, including related information theoretic accounts. |
| **Order of Authors:** | Anne-Marike Schiffer |
| | Christiane Ahlheim |
| | Moritz F Wurm |
| | Ricarda I Schubotz |
| **Suggested Reviewers:** | Stefan J Kiebel, Dr<br>Max Planck Institute for Human Cognitive and Brain Sciences<br>kiebel@cbs.mpg.de |
| | Emrah Duzel, Professor<br>University College London<br>e.duzel@ucl.ac.uk |
| **Opposed Reviewers:** | |

## Title

# Surprised at all the Entropy: Hippocampal, Caudate and Midbrain Contributions to Learning from Prediction Errors

## Authors and Affiliations

Anne-Marike Schiffer* (1), Christiane Ahlheim (1,2), Moritz F. Wurm (1), & Ricarda I. Schubotz(2,1)

1Motor Cognition Group, Max Planck Institute for Neurological Research, Cologne, Germany

2Westfaelische Wilhelms-Universitaet Muenster, Institut fuer Psychologie, Muenster, Germany

*Corresponding Author

schiffer@nf.mpg.de

## Abstract

Influential concepts in neuroscientific research cast the brain a predictive mechine that revises its predictions when they are violated by sensory input. While this is famously implemented in the predictive coding account of perception, it also relates to learning. Learning from prediction errors however has been suggested for the hippocampal memory system as well as for the basal ganglia. The present fMRI study used an action-observation paradigm to investigate the contributions of the hippocampus, caudate nucleus and midbrain dopaminergic system to different types of learning: learning in the absence of predicton errors, learning from prediction errors, and responding to the accumulation of prediction errors in unpredictable stimulus configurations. We conducted analyses of the regions of interests' BOLD response towards these different types of learning, implementing a bootstrapping procedure to correct for false positives. We found both, caudate nucleus and the hippocampus to be activated by perceptual prediction errors. The hippocampal responses seemed to relate to the associative mismatch between a stored representation and new sensory input. Moreover, its response was significantly influenced by the average information, or entropy of the stimulus material. Behavioural measures indicated, that memory for information received under increasing entropy is worse than memory for stable representations. The habenula mirrored the caudate's responses to perceptual prediction errors unrelated to reward. Lastly, we found that the substantia nigra diplays a disparate response pattern: it was activated by the novelty of sensory input. In sum, we established differential involvement of the hippocampus, caudate nucleus and midbrain dopaminergic system in different types of learning. We relate learning from perceptual prediction errors to the concept of predictive coding, including related information theoretic accounts.

## Introduction

The notion of the brain as a predictive machine pervades contemporary neuroscientific concepts [1-6]. One great achievement of the approach is that it brings perception and learning into the proximity [7]. If the brain constantly predicts its sensory input [8-9] it has to learn correct models of its environment to achieve these predictions [10]. This idea delivers powerful accounts to explain cortical responses [11], especially in primary sensory cortices [9] and the cortical motor network [12]. The contributions of subcortical and allocortical components, however, may not have received due attention. The present study investigates how the caudate nucleus and hippocampus may contribute to learning in a predictive framework.

The predictions of the brain and the update mechanisms of these predictions are encompassed in the predictive coding account of perception [2, 10, 13, 14]. This account recasts the brain as a Bayesian inference machine [15]. Perception thus relies on probabilistic models at each level of cortical hierarchy [8, 9, 11, 16]. Each of these models equals a representation and the probability of sensory input at the level below [8, 11, 13, 14], given the representation accords to the most likely state of the environment. The model sends these predictions of probable lower level activity via backward projections to the level below [2, 11]. If the sensory input at this lower level matches the predictions, the signal is filtered [9, 11, 13]. If the sensory input does not match the predictions, the difference is signaled via forward connections to the next higher level [11]. This difference is called the prediction error [8, 11]. It could also be described as the surprise at the sensory input [8, 17-19]. The prediction errors cause an adjustment of the model at the higher level. This adjustment can pertain to learning if the probabilities encompassed in the model and thus its predictions are altered as a result of the prediction errors [8], or if the internal model is replaced by a model that contributes more precise predictions [20]. Each inferential perception can thus potentially bring about learning [8,10]. What type of learning occurs depends on the reliability of information. If prediction errors accumulate, the environment is said to contain a lot of entropy [8, 18,19]. In psychological terms, entropy can be translated to uncertainty [21]. Volatility, another measure of uncertainty, has been shown to influence learning rate [22].

Neuroscientific research on learning has discussed the interplay and competition of two learning systems to a large extent [23-27]. One of these systems relies on the striatum, while the other is understood to be hippocampus-based. Both systems have been associated with learning from violated predictions [28-31]. Moreover, both systems receive projections from the midbrain dopamingergic system that seems to be involved in each systems' respective learning mechanisms [28, 31 – 33].

The hippocampal memory system is regarded as an associative mismatch detector [29-34], which it is responsive when the predictions of stored representations are violated by events that were previously not associated with the stored representation [24, 35]. Importantly, the hippocampus and its underlying dopaminergic projections have been proposed to underlie sequential learning [33, 35 – 37] and code for violations of sequences [29, 38] Lastly, new results have suggested that hippocampus is not responsive to novelty or violated predictions per se, but uncertainty [18], signifying the learning that oddballs can occur

The striatum and its underlying dopaminergic projections have famously been established to be responsive to prediction errors in reward context [30,31], a finding that has been confirmed in humans [39,40]. Moreover, recent imaging studies suggest that perceptual prediction errors, i.e. violated expectations unrelated to reward, also activate the striatum [41 - 43].

The current study aimed to dissociate the contributions of the hippocampal and striatal systems to different types of learning that are marked to different degrees by novelty and prediction errors. The first type of learning that was investigated was the acquisition of new representations that we call internal models (*new originals*, hereafter). This activity change basically pertains to the adaptation of novelty responses, signified by an attenuation of the BOLD response. The second type of learning we investigated was the adaptation of predictions when the expectations of a model were violated by a divergent version (*divergents,* hereafter) that was thereafter repeated. The third type of learning was the response to constant violation of a model by unpredictable versions (*unpredictables*, hereafter). This manipulation did not allow predicting the content of a model, corresponding to a type of learning that is signified by a lot of uncertainty.

We hypothesized that the hippocampal memory system should be activated to a larger extend by associative novelty than novelty per se and thus show more activity towards the unpredictable movies and divergent movies than the novel movies. Moreover, we expected the hippocampus to be responsive to the entropy that resulted from repeated violations of model predictions [18].

With regard to striatal responses during learning, we focused on a subdivision of the caudate nucleus that has previously been associated with prediction errors [41] and expected this part of the striatum to be only responsive to prediction errors and not to respond to novelty. We therefore predicted that this caudate nucleus subdivision should decrease during repeated presentation of the same divergent model. We also predicted that this area should be activated more by the unpredictable movies that pose an accumulation of prediction errors than by the divergent movies. Lastly, with regard to the midbrain dopaminergic system, we predicted firstly, that the habenula, which has previously been associated with prediction errors and uncertainty, would mirror the caudate response. Secondly, we investigated exploratively whether the substantia nigra, the input to caudate and hippocampus would yield activity in line with one or both structures, or would show a separate response pattern.

## Results

1. Behavioural results

The correlation between the sum of recalled actions per condition and the exposition frequency was significant ( r= .458, p < .001). The repeated measures ANOVA on the standardized residuals of the number of recalled actions and the factor CONDITION yielded significance ($F_{(2.81,50.54)}$= 3.505, p = .024; Greenhouse Geisser corrected for non-sphericity). Actions in the condition *divergents* were named more often (mean = 2.26 times), than actions in the conditions *new originals* (0.47 times) and *unpredictables* (1.53 times).

2. ROI analyses

2. 1. Contrast relating to acquisition

There was a significant attenuation of activity with repeated exposures of the *new originals* in the hippocampus ($t$ = 1.45; $t_{crit\ 5\%}$ = 0.65; p < .05). The substantia nigra showed attenuation of activity in the same parametric contrast ($t$ = 2.00, $t_{crit\ 5\%}$ = 1.56, p < .05). There was no significant attenuation of activity in the caudate nucleus ROI.

The substantia nigra was the only structure that showed a main effect for the processing of *new originals* vs. *divergents*. It was significantly less activated by the processing of *divergents* compared to *new originals* ($t$ = -1.225; $t_{crit\ 5\ \%}$ = -1.225; p < .05).

2.2. Contrasts relating to adaptation:

The hippocampal ROI revealed a significant attenuation of activity with the repeated exposure of *divergents* ($t$ = 0.88; $t_{crit\ 5\%}$ = 0.62; p < .05).

2. 3. Contrasts relating to unpredictability:

Processing of the unpredictables activated the hippocampal ROI significantly more than processing of *new originals* and *singletons* ($t$ = 1.65, $t_{crit\ 5\%}$ = 1.60, p < 0.05). In the caudate ROI there was more activity for the processing of *unpredictables* at all stages than for the processing of *divergents* ($t$ = 2.33; $t_{crit\ 5\%}$ = 1.76, p < 0.05). The habenula ($t$ = 2.58; $t_{crit\ 5\%}$ = 1.79; p < 0.05) and the substantia nigra ($t$ = 2.45; $t_{crit\ 5\%}$ = 1.56; p < 0.05) were similarly activated more by *unpredictables* than by *divergents*. There was no attenuation with the repeated exposure of *unpredictables* in any ROI at p< 0.05.

We conducted a repeated measures ANOVA testing for main effects of CONDITION (*new originals, divergents, unpredictables*) on attenuation effects in the hippocampal ROI. The repeated measures ANOVA yielded a significant main effect of CONDITION. This effect was due to significant differences between *new originals* and *unpredictables*. Since the dependent variable reflected the slope of the attenuation, this results indicate an interaction between the course of the ATTENUATION and the CONDITION in the hippocampus.

2.4. Bayesian modelling:

The hippocampal ROI yielded significant activity for the modelling of Bayes’ian entropy in the *unpredictables* ($t = 1.83$; $t_{crit5\%} = 1.59$, $p < 0.05$). Activity in the caudate ROI for the modelling of Bayesian surprise during the observation of *unpredictables* approached significance at p = .054 ($t = 2.98$;$t_{crit\,5\%} = 3.00$; $t_{crit\,10\,\%} = 2.73$, $p < 0.1$)

## Discussion

The hippocampal ROI showed no significant difference between the acquisition of a new model and the adaptation to a changed model. In fact, both processes were signified by a decrease in activity as hypothesized. Interestingly, we found significantly higher activity for the unpredictable change to a known model than for complete novelty, in line with the associative mismatch account [29]. Lastly, the hippocampus showed an activity increase over the course of *unpredictables* that reflects the Shannon entropy, or average surprise, elicited by the prediction errors inherent in this condition (Figure 4).

In contrast to the response pattern observed for the hippocampal ROI, the caudate ROI was significantly more activated by the processing of the prediction error profuse *unpredictables* than by the processing of the eventually predictable *divergents*. Descriptively, the caudate ROI also showed a trend towards activity corresponding to Bayesian surprise (Figure 5).

Finally, as predicted, the habenula reflected the caudate response to the occurence of prediction errors in the *unlearnables* The substantia nigra displayed novelty responses. These were in contrast to the activity pattern revealed by the hippocampus not dependent on associative novelty but reflected novelty per se.

Predictive Coding and the Hippocampus

The hippocampus showed a clear involvement in the acquisition of a new model and adaptation of an old model to change. It decreased in activity with each iteration of the *new originals* and *divergents* that was observed.

Activity decrease as a hallmark of learning has found multiple implementations in different predictive paradigms (see [44] for a recent review). Predictive or inferential accounts of brain function explain why a decrease in activity can be regarded a sign of learning [2, 9, 11, 14]. To resurrect the picture, the brain builds models of likely perceptions [11, 14, 45, 46]. Sensory input is predicted on the basis of these internal models. The model effectively filters all anticipated information and thus modulates cortical activity to represent only surprising, informative input [2, 11]. This activity, due to prediction errors, can either cause the model to loose weight in predicting the sensory input (and thus effectively being replaced by another model, cf. [20]), or the change of the models' predictions [10] pertaining to learning. Decrease of neural activity over repeated iterations of a model are therefore regarded as a sign of learning [46, 47]. As the model gets better, there are less prediction errors, causing less cortical activity. The fact that the model gets more precise in predicting, and thereby filtering input, means it has learnt.

Predictive coding is usually regarded to deal with current, not anticipated sensory input [12, 5]. However, viewing hippocampal activity from a predictive coding perspective reveals how predictions into the near future could be mediated. Combining sensorimotor cortical responses as explained by predictive coding [2,11] with models of hippocampal function [36, 38] explains how predictions of consecutive events can be established and matched with sensory reality. Two functions of the hippocampus relate to this account: first of all the hippocampus is regarded to store compressed representations of cortical activity [24, 47, 48]. Secondly, it has the capability for coding sequential events [47, 49 – 51], for example in spatial navigation [23, 29, 34, 52, but see 53] and learning of episodes [49, 50]. These functions relate to a relational representation [50, 51, 54, 55, but see 56], i.e., a sparse coding of cortical patterns, and importantly also their relation in time and space. This relation is achieved by small overlaps between the sparse representations of the cortical patterns [57].

Prediction of sequential events [47, 50] and spatial navigation [58] relies on the succession of cortical patterns [59], coding the (visual) input at a given time, and the (visual) input that should follow. To predict the next pattern in the sequence, the hippocampus can use the above-mentioned minimal overlap between the cortical representations. Importantly, hippocampal representations can be back-projected to the cortex, the putative mechanism behind retrieval and implicit learning [36]. If the overlap is reinstalled by repeatedly experiencing the sequence of cortical patterns, it is strengthened [56, 57]. The predictive coding account suggests that cortical patterns are diminished once they are predicted. If one cortical pattern that is part of a condensed sequential representation was elicited by unpredicted (e.g. visual) input, this would lead to a retrieval of the stored representation (cf. pattern completion, [29, 38]) that predicts the next cortical pattern in the sequence [29, 59]. If this cortical pattern occurred, it would be effectively filtered according to the predictive coding account [44]. This less in cortical activity may cause comparatively less encoding or weight change in the hippocampus, compared to a perception that does not fit the predicted input; this account explains novelty signals and especially signals reflecting the mismatch between predictions and sensory input as unfiltered prediction errors.

We could show that long stimulus sequences, i.e., actions, that are new to the observer lead to a stepwise decrease in hippocampal activity. We propose that the sequence of actions in the scripts became predictable and the associated sequence of cortical patterns resulted in a filtering of the sensory input. The decrease in hippocampal activity can therefore be understood as a sign of an increasingly valid model that predicts the course of action [44, 45]. It is important to note that the predictions of sensory input entailed conceptual predictions, as the different shots of each script negated surface-similarities.

The associative mismatch account of hippocampal function [29] in fact captures the same elements as predictive coding. It predicts that anticipated input will result in lower activity than unpredicted input. Moreover, Kumaran and Maguire [29] could show that unpredicted input also elicits more activity than novel input. Thus, not novelty, but the mismatch between expected and perceived sequences activated the hippocampus [29]. This finding coined the term of "associative mismatch detector" to functionally describe the hippocampus proper. The present study extends this notion in an important manner. The unpredictable courses of known movies elicited more activity than completely new movies. The finding that novel items (*singletons* and 1st *new originals*) elicit less activity than *unpredictables* that relate to a previous association can also be recast in terms of predictive coding. As described previously, predictive coding rests on Bayesian inference. That is, the first of frequently paired items starts to predict the second item with a high conditional probability. If this pairing is consistent, the brain experiences little entropy and will therefore not expect any deviations. A violation of this prediction results in a higher activation than the encounter of an action movie that is not encompassed in a recently acquired internal model, as in the case of the first new originals and singletons. If no solid internal model exists so far, the input will be filtered only to the degree that is proposed by known action semantics. In comparison to the episodic internal model trained for the *unpredictables*, the internal model for the *new originals* does not ascribe a solid probability to specific episodically acquired predictions. Thus, the mismatch signal is smaller for the more lenient semantic predictions.

### Entropy in the hippocampus

The current results suggest that the hippocampal activity reflects Shannon entropy of the unpredictable courses (cf. [19]). Shannon entropy mirrors the average surprise within a stimulus stream [18]. In psychological terms we can therefore regard entropy as a measure of uncertainty concerning predictions. While the responsiveness of the hippocampus to Shannon entropy replicates a result by Strange and colleagues [18], it also expands our knowledge on hippocampal function substantially. The experiment by Strange and colleagues [18] dealt with learning of statistical regularities. It did however not allow learning to predict the next item, but only learning to predict the rate of occurrence of items [18, 60]. On the other hand, a related study by Harrison and colleagues [60] investigated the involvement of the hippocampus in learning the likelihood of a transition between two successive items. These authors found no indication of hippocampal coding for entropy [60]. In the current study the hippocampus was sensitive to the entropy caused by unpredicted sequences of actions, thus indicating that the hippocampus is sensitive to the predictability of transitions in very complex stimuli, and without a priori knowledge of all transitions or stimuli that will occur. This latter fact seems to be relevant when considering that Strange and coworkers [18] have suggested that the hippocampus does not encode the stimuli that violate predictions, but the fact that these occur. However, the stimuli used by Strange and colleagues [18] were all a priori known. Thus their would have needed no encoding. However, what could be learnt was an expectation of their probability, which in fact equals entropy. However, the current study employed action movies and violations stemmed from previously unassociated actions within the sequence. If these actions had not been encoded, future violations and the entire unpredictability could not have been detected. In fact, if the content of violation had not been encoded at all, the responses towards the *unpredictables* would have mirrored the responses towards the *divergents*.

Having said that, it is interesting that the free recall rates for *divergents* surpassed that for *unpredictables*, suggesting a less successful encoding of the *unpredictables*. This finding maybe not surprising, given the fact that *unpredictables* did not possess the reliability to enable future valid predictions. We thus find tentative evidence that while stimulus sequences exposing high Shannon entropy are encoded to a certain degree, the encoding is not as successful as that for low-entropy or stable sequences.

Based on the results of the present study, we propose that the hippocampus adapts its models of sequential sensory input as implied by the associative mismatch account [29]. Thus, its activity is different from that of the putatively underlying dopaminergic projections from the substantia nigra, that are sensitive to novelty, but not associative novelty, i.e., associative mismatch (but see [61]). Moreover is the hippocampus sensitive to the uncertainty under which it receives information and encodes the uncertainty-eliciting input to a specific degree.

The caudate nucleus in perceptual prediction errors

The caudate nucleus showed a higher response to *unpredictables* than to *divergents*. Each *unpredictable* contained a breach of expectation on the content level, that is the sequence of actions. But only the first *divergent* contained a breach of expectation on the content level while each subsequent *divergent* version of the same movie repeated the same diversion from the original script. On a higher level of description, each breach of expectation of the *unpredictables* that occurred after the second iteration was fully predictable as such, (albeit not predictable with regard to the post-preach content). Caudate nucleus activity was therefore driven by prediction errors on the content level, indicating a lack of meta-learning. Caudate signaling of prediction errors is noteworthy in itself, as only few studies have discussed the striatal involvement in non-reward related prediction errors [41-43]. The dominant account for striatal functioning is the *temporal difference model* that is usually associated with reward related learning [3,31]. Only one recent study has applied prediction errors in terms of predictive coding to striatal function [42]. The results of the present study therefore contribute substantially to a new understanding of striatal signaling: the indication of prediction errors on a perceptual level, irrespective the presence of reward or punishment. On a related note, it is interesting that the habenula mirrored the caudate activity. This result substantiates our previous finding [41] of the habenula's involvement in coding for perceptual prediction errors. This result and its replication are highly interesting, as the habenula is generally understood to code for punishing or „worse than expected“ outcomes [62]. In close keeping with an argument put forward by Friston and colleagues [63] prediction errors can concern the valence of an outcome. However, the involvement of the habenula in perceptual prediction errors could indicate that prediction errors as an outcome of a predictive process can have a valence themselves, possibly motivating the improvement of internal models.

## Materials and Methods

2.1 Subjects:

19 right-handed, healthy participants (7 women, age 22-30 years; mean age 25.3 years) took part in the study. The participants were right handed as assessed with the Edinburgh Handedness Inventory [64]. The experiment was approved by the local ethics committee of the University of Cologne and in accordance with the Declaration of Helsinki. All participants were health screened by a physician and gave written informed consent.

2.2 Stimuli and Task:

The stimulus material contained 37 different movies of 8 to 12 seconds length (mean 9.2 sec; standard deviation 1.39 sec). The movies were shot from the third-person perspective, not showing the actor's face. They contained every-day actions taking place at a table. Most movie scripts, e.g. making a sandwich, existed in 2 versions (*divergents*). Some movie scripts existed in 6 different versions (*unpredictables*). All of these scripts had an identical beginning, but started to diverge at some individual point, whereafter no commonality existed (Figure 1). Each version of the divergents (a and b) was filmed 18 times. Of the six-versions scripts, version a and b were shot 9 times each, whereas versions c, d, e and f were shot only once. Thus, even though the same script appeared repeatedly during the pre-experimental exposition (see below) and for the movies that returned in their original version (*originals*, herafter), as well as for the *divergents* and *unpredictables* also during the experiment, the exact same shot of each script occurred only once during the pre-experimental and the experimental session. This method was employed to minimize surface-similarities between the movies and avoid surface-reference perceptual priming.

The experiment consisted of a pre-experimental exposition of the movie material and an fMRI session starting 15 minutes after the end of the pre-exposition. During the pre-experimental exposition session, participants were seated in a sound-attenuated chamber facing a computer screen. Distance to the screen was adjusted to ensure that the video displayed on the screen did not extend 5° of visual angle. The participants watched 27 scripts, a third of which was displayed three times, another third six times and the last third nine times, but in a randomized fashion over the course of the 28 minutes lasting session. As mentioned above, the participants watched one version of each script; but each repetition was another shot of the same script (minimal distance 4 different scripts in between). Questions concerning whether some action or another was part of the immediately preceding script (e.g. “grasping an apple?”) were posed on average after every fifth script (mininimum one movie, maximum 11 movies in between, standard deviation 2.1) to ensure ongoing attention to the stimulus material. Participants received visual feedback for 400 ms on whether they had answered correctly, incorrectly, or too late. After pre-exposition, participants were transferred directly to the fMRI chamber.

### 2.3 FMRI session

The fMRI session encompassed display of 36 different scripts. Each script was repeated over the experiment. Nine scripts that had previously been displayed during the pre-exposition returned nine times in the fMRI session in the same version as before (*originals*). Another nine of the pre-experimentally shown scripts were presented nine times in the fMRI session only in their complementary version (*divergents*). The last nine scripts appeared in five different versions during the fMRI, each being displayed only once (*unpredictables*). One third of all movies (including the *originals*, the *divergents* and the *unpredictables*) had previously been displayed three times each, another third six times each, and one third nine times each. The design moreover encompassed three scripts that were repeated nine times during the fMRI session and completely new to the participants at first exposure (*new originals*, hereafter). Finally, there were six single movies that were displayed only once and had not been pre-exposed previously (*singletons*, hereafter) (Table 1).

Immediately after the fMRI session, participants filled in a questionnaire encompassing a free-recall task for the movie scripts.

### 2.4 Data Acquisition

The functional imaging session took place in a 3T Siemens Magnetom Trio scanner (Siemens, Erlangen, Germany). In a separate session, prior to the functional MRI, high-resolution 3D T-1 weighted whole-brain MDEFT sequences were recorded for every participant (128 slices, field of view 256 mm, 256 by 256 pixel matrix, thickness 1 mm, spacing 0.25 mm)

The functional session engaged a single-shot gradient echo-planar imaging (EPI) sequence sensitive to blood oxygen level dependent (BOLD) contrast (28 slices, 4 mm thickness, 0.6 mm spacing; in-plane resolution of 3 x 3 mm) parallel to the bicommisural plane, echo time 30 ms, flip angle 90°; repetition time 2000 ms; serial recording). Following the functional session immediately, a set of T1-weighted 2D-FLASH images was acquired for each participant (28 slices, field of view 200 mm, 128 by 128 pixel matrix, thickness 4 mm, spacing 0.6 mm, in-plane resolution 3 by 3 mm).

### 2.5 FMRI Data Analysis

Functional data were offline motion-corrected using the Siemens motion protocol PACE (Siemens, Erlangen, Germany). Further processing was conducted with the LIPSIA software package [65]. Cubic-spline interpolation was used to correct for the temporal offset between the slices acquired in one scan. To remove low-frequency signal changes and baseline drifts, a highpass filter was applied. The filter length was adapted to the rate of occurrence of the rarest event and was different for all analyses containing *new originals* compared to the other analyses. The filter in the contrasts investigating only *unpredictables* and *divergents* was set at 1/85 Hz. The (parametric) contrasts containing new originals were highpass filtered at 1/90 Hz. The matching parameters (6 degrees of freedom: 3 rotational, 3 translational) of the T1-weighted 2D-FLASH data onto the individual 3D MDEFT reference set were used to calculate the transformation matrices for linear registration. These matrices were subsequently normalized to the standardized Talairach brain size (x = 135 mm, y = 175mm, z = 120mm [66]) by linear scaling. The normalized transformation matrices were then applied to the functional slices, to transform them using trilinear interpolation and align them with the 3D reference set in the stereotactic coordinate system. The generated output had thus a spatial resolution of 3 by 3 by 3 mm. A spatial Gaussian filter of 5 mm FWHM was applied.

The statistical evaluation was based on a least-square estimation using the general linear model (GLM) for serially auto-correlated observations [67]. Temporal Gaussian smoothing (4 seconds FWHM) was applied to deal with temporal autocorrelation and determine the degrees of freedom [67].

The design matrix was generated by hemodynamic modeling using a γ-function and its first derivate. The onset vectors in the design matrices were modeled in a time-locked event-related fashion and set to the point in time (hereupon 'breach') when the movie (in the conditions *divergents* and *unpredictables*) differed from its original pre-experimental exposition version. The originals and new originals were modeled after the point in the movie that would have been the breach, if they had been displayed in their complementary version. This pseudo post-breach modeling was employed for the originals and new originals, as all scripts were counterbalanced in their assignment to conditions across participants. Thus some participants could have encountered in the function of divergent what to others was the original, or even new original. We thus ensured that the measured effects did not stem from the identity of scripts or comparative length, but solely their assigned condition in the experiment. The breach had previously been visually timed to the moment when movement trajectories revealed that either the manipulation or the reach-for-object was different from that in the originals. The length of the modeled events corresponded to the length of the script from the breach to the end of the script (mean: 6.57 sec; STD: 1.78 sec).

### 2.5.1. Region of interest (ROI) definition

We used the 3D T1-weighted whole-brain scans of each participant to individually segment four ROIs: left and right caudate nucleus (Figure 1), left and right habenula, left and right substantia nigra (Figure 3) and left and right hippocampus proper (Figure 2). The habenula, substantia nigra and hippocampus ROIs were delimited according to anatomical landmarks. The caudate ROI was created using the coordinates of the peak voxels activated for violated predictions in a previous study [41] and choosing a radius of 4 voxels. The resulting 3-D area was then clipped in each brain individually to exclude the internal capsule and ventricles. In 3 participants, clipping the caudate ROIs to exclude the ventricles and internal capsule left nothing of the caudate ROI remaining. These participants were therefore excluded from the analysis.

The fMRI data analysis proceeded in two steps. In a first step, we modeled each condition of interest (*divergents, unpredictables* and *new originals*) parametrically. Therefore, we generated three separate design matrices, each containing three event types, two times the movie type of interest and null events. For example, the design matrix for *unpredictables* contained as a first event type all *unpredictables* with an amplitude vector of one. As a second event type, it contained all *unpredictables* with an amplitude vector corresponding to the specific script's iteration in the fMRI session. (the first iteration of one script was assigned an amplitude of five, the second an amplitude of four, and so forth). The last event type in the design matrix were null events, assigned an amplitude vector of one. The same set up applies to the design matrices for the parametric attenuation modeling of *divergents* and *new originals*. In a second step, we contrasted the *unpredictables* with the *divergents* and the *divergents* with the *new originals* to investigate the relative and persistent involvement of the hippocampus proper and the striatum, i.e. caudate nucleus, in the processing of the different movie types. Thus, the fourth design matrix contained as the first event-type all *unpredictables*, each with a vector amplitude of one, as the second event-type all *divergents*, with a vector amplitude of one and lastly as a third event-type all null-events with a vector amplitude of one. The fifth design matrix contained the event-types *divergents*, *new originals* and null-events, all modeled with a vector amplitude of one. The sixth analysis contrasted 12 randomly chosen *unpredictables* (each with an amplitude vector of one) with the first presentation of the *new originals* and *singletons* (with the same ampitude) and also contained null-events.

2.5.2. Bayesian modeling analysis

We calculated the responses of all four bilateral ROIS to the Shannon entropy (Figure 5) and surprise (Figure 4) ascribed to the content-development of the *unpredictables*. We assumed that the brain should behave like an ideal observer and hence ascribe the probability of an item according to:

$$p(x_i) = \frac{n^i_j + 1}{\sum_k n^k_i + 1} \; .$$

This model is in close keeping with the approach taken by Strange and coworkers [18]. The n signifies the total number of occurrences of a movie version so far. In the numerator, the number concerns the observation of the exact version of the movie per se, in the denominator it concerns the observations of all other movie versions so far. The addition of the value 1 shape a Dirichlet distribution, that accords to an ideal observer.

Following previous approaches [17-19], surprise at an outcome was calculated as:

$$I(x_i) = -\ln p(x_i) .$$

This term, also known as the 'negative evidence', indicates the amount of information that is conveyed by the observation [8].

Another important construct that describes the influence of observations is Shannon entropy. Shannon entropy is again a term derived from information theory [19] (but see [21]) and describes the average surprise in a series of observations ([17]. Shannon entropy is therefore mathematically calculated as:

$$H(X) = \sum_j -p(x_i) \ln p(x_i)$$

[17-18, 21]. The negative probability multiplied with the logarithmic probability of each event i is summed for all events that could have occurred within one trial j. (We employed the natural logarithm, but binary approaches have been used (cf. [18]). If all observations are equally likely and appear equally often, each event is surprising, as it cannot be predicted [17]. This is the setup of the highest Shannon entropy. If Shannon entropy is large, each event is very informative [8, 17, 19].

The second level analysis employed a permutation analysis to correct for false-positives [68]. For each of the 8 contrasts, we calculated 2000 different one-sample *t* tests for each of the four ROIs. The important manipulation consisted in a different reversal of experimental and control condition in one to 16 subjects in all 2000 *t* tests. It can thus be determined, whether the analysis that agrees with the experimental setup in all participants reaches a higher *t*-value than randomly permuted analyses. This would then indicate, that the activity revealed in the contrast is best accounted for by the contrast between experimental and control condition and not due to noise. The benefit of such a boot-strapping approach is that the *t* tests do not assume a Gaussian distribution, but calculate the distribution based on the the variance in the data [68]. This is important, as the use of a Gaussian distribution does not necessarily fit activity in a spatially circumscribed ROI. The cut-off $t$ ($t_{crit}$) for significance testing was set at p = .05. This means that 1900 permutations of the assignment between subjects and conditions must result in a lower $t$ than the original experimental assignment wherein the control condition is used as control condition and the experimental condition used as experimental condition for all 16 subjects.

### 2.6 Behavioural data analysis

After the fMRI session, participants were asked to recall as many actions as they could remember. To test if the different actions were differently well remembered depending on their condition, these free recall rates were analyzed. Therefore, it was counted how many movies of each condition were recalled by each subject and how often each of the recalled movies had been seen during the experiment (pre-exposition and functional scanning). Note that it was aggregated for each version of the movies, i.e. divergent movies have been exposed 3+9, 6+9, or 9+9 times, whereas all new originals had been exposed 9 times (during the functional scanning). The average number of expositions was calculated by summing up the exposition rates of all movies and dividing it by the number of recalled movies of the condition. The inferential analysis was performed in three steps.

At first, the influence of the exposition frequency was partialed out by running a multiple regression with the sum of the recalled actions (per condition) as dependent and the number of pre-expositions as independent variables. The standardized residuals of this analysis, i.e., the information that was not explained by exposition frequency, served as dependent variable in the analysis of the condition effect. To that end, a repeated-measures ANOVA was calculated with the factor CONDITION (*originals, new originals, divergents, unpredictables, singletons*).

It must be borne in mind that all unpredictable versions of one movie shared common actions in the common beginning of the script. Moreover, the objects in different versions were sometimes the same as in other versions, while the manipulation of the object differed. For instance, all 6 different versions of one particular movie (the pre-exposed version as well as the five unpredictable versions during the fMRI) contained a piggy bank. Naming a script from the *unpredictables* condition was therefore not necessarily harder than naming a script from the *originals, new originals* or *divergents* condition.

## Acknowledgments

We wish to express our gratitude to Alexander Wagner for his advice on the Bayesian modeling approach and Kirstin Ulrichs for her substantial contributions to stimulus material preparation and data collection.

## References

1. Bubic A, von Cramon DY, Schubotz RI (2010) Prediction, cognition and the brain. Frontiers in Human Neuroscience. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2904053/ accessed: 12/22/11.
2. Huang Y, Rao RPN (2011) Predictive coding. Wiley Interdisciplinary Reviews: Cognitive Science 2: 580–593.
3. Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mescencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16(5): 1936-1947.
4. Schubotz RI (2007) Prediction of external events with our motor system: towards a new framework. TiCS 11(5): 211-218.
5. Schütz-Bosbach S, Prinz W (2007) Prospective coding in event representation. Cognitive processing 8(2): 93-102.
6. Wolpert DM, Flanagan JR (2001) Motor prediction. Current Biology 11(18): R729-R732.
7. Fiser J, Berkes P, Orbán G, Lengyel M (2010) Statistically optimal perception and learning: from behavior to neural representations. TiCS 14: 119-130.
8. Friston K (2010) The free-energy principle: a unified brain theory? Nat. Reviews. Neurosci, 11(2): 127-138.
9. Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2: 79–87.
10. Friston KJ (2002) Dunctional integration and inference in the brain. Prog. Neurobiol. 68(2): 113/143
11. Friston KJ (2005) A theory of cortical responses. Philos. Trans. R. Soc. Lond., 360: 815–836.

12. Kilner JM Friston KJ, Frith CD (2007) Predictive coding: an account of the mirror neuron system. Cogn Process. 8(3):159-66.

13. Kersten D, Mamassian M, Yuille A (2011) Object Perception as Bayesian Inference. UC Los Angeles: Department of Statistics, UCLA. Available at: http://escholarship.org/uc/item/69d797cq. accessed:12/22/11.

14. Knill DC, Pouget A (2004) The Bayesian brain: The role of uncertainty in neural coding and computation for perception and action, TiN 27(12): 712-719.

15. Crapse TB, Sommer MA (2008) The frontal eye field as a prediction map. Progress in brain research 71(08): 383-90.

16. Kiebel SJ, Daunizeau J, Friston KJ (2008) A Hierarchy of Time-Scales and the Brain. PLoS Comput Biol 4(11): e1000209. Available at: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000209. Accessed 12/22/11

17. Doya K, Ishii S (2007) A probability primer. In: Doya K, Ishii S, Pouget A, Rao R, editors. Bayesian Brain: Probabilistic Approach to Neural Coding and Learning. Cambridge: MIT Press. pp. 3-13.

18. Strange BA, Duggins A, Penny W, Dolan RJ, Friston KJ (2005) Information theory, novelty and hippocampal responses: unpredicted or unpredictable? Neural Netw. 18: 225-230.

19. Shannon, CE (1948) A mathematical theory of communication. Bell System Technical Journal 27: 379-423 and 623-656.

20. Wolpert DM, Kawato M (1998) Multiple paired forward and inverse models for motor control. Neural Networks 11(7-8):1317-1329

21. Luce RD (2003) Whatever happened to information theory in psychology? Review of General Psychology 7(2): 183-188.

22. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) Learning the value of information in an uncertain world. Nat Neurosci, 10(9): 1214-1221.

23. Doeller CF, King JA, Burgess N (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. PNAS 105(15): 5915-5920

24. Atallah HE, Frank MJ, O'Reilly RC (2004) Hippocampus, cortex, and basal ganglia: Insights from computational models of complementary learning systems. Neurobiology of Learning and Memory 82: 253–267

25. Poldrack RA, Packard MG (2003) Competition among multiple memory systems: converging evidence from animal and human brain studies. Neuropsychologia 41(3): 245–251.

26. Packard MG, Knowlton BJ (2002) Learning and Memory Functions of the Basal Ganglia. Annu. Rev. Neurosci. 25:563–93.

27. Ashby FG, Alfonso-Reese LA, Turken AU, Waldron EM (1998) A Neuropsychological Theory of Multiple Systems in Category Learning. Psychological Review 105(3): 442-481

28. Shohamy D, & Adcock RA (2010) Dopamine and adaptive memory. TiCS 14(10): 464-472.

29. Kumaran D, Maguire EA (2006) An unexpected sequence of events: mismatch detection in the human hippocampus. *PLoS biology*, *4*(12), e424. doi:10.1371/journal.pbio.0040424 Available at: http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0040424. Accessed 12/22/11.

30. Schultz W (2000) Multiple reward signals in the brain. Nature Rev Neurosci 1(3): 199-207.

31. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275(5306): 1593-9.

32. Graybiel AM (2005) The basal ganglia: learning new tricks and loving it. Current Opinion in Neurobiology 15(6): 638-644.

33. Jay TM (2003) Dopamine: a potential substrate for synaptic plasticity and memory mechanisms. Progress in Neurobiology 69(6): 375-390.

34. Duzel E, Habib R, Rotte M, Guderian S, Tulving E, et al. (2003) Human hippocampal and parahippocampal activity during visual associative recognition memory for spatial and nonspatial stimulus configurations. J Neurosci 23: 9439–9444.

35. Lisman JE, Grace AA (2005). The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. Neuron 46(5): 703-713

36. Gluck MA, Myers C, Meeter M (2005) Cortico-hippocampal interaction and adaptive stimulus representation: A neurocomputational theory of associative learning and memory. Neural Networks 18: 1265 – 1279.

37. Kumaran D, Duzel E (2008) The Hippocampus and Dopaminergic Midbrain: Old Couple, New Insights. Neuron, 60(2):197-200.

38. O'Reilly R, Norman KA (2002) Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. TICS 6(12) 505 -510.

39. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston KJ, Dolan RJ (2004) Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. Science, 304(5669), 452-454.

40. O'Doherty J, Dayan P, Schultz J. Deichmann R, Friston KJ, Dolan RJ (2003) Reward representations and reward-related learning in the human brain: insights from neuroimaging. Current Opinion in Neurobiology 14(6): 769-776.

41. Schiffer A-M, Schubotz RI (2011) Caudate nucleus signals for breaches of expectation in a movement observation paradigm. Frontiers in Human Neuroscience 5. Available at http://www.frontiersin.org/human_neuroscience/10.3389/fnhum.2011.00038/full. accessed 12/22/11.

42. den Ouden HEJ, Danizeau J, Roiser J, Friston KJ, Stephan KE (2010) Striatal Prediction Error Modulates Cortical Coupling. J. Neurosci. 30(33):11177-11187

43. Spicer J, Galván A, Hare TA, Voss H, Glover G, Casey BJ (2007) Sensitivity of the nucleus accumbens to violations in expectation of reward. Neuroimage 34: 455-461.

44. Colder B. (2011). Emulation as an Integrating Principle for Cognition. Frontiers in Human Neuroscience. Available at: http://www.frontiersin.org/human_neuroscience/10.3389/fnhum.2011.00054/full. accessed: 12/22/11.

45. Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T (2008) Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci.1 (9), 1004-1006.

46. Summerfield C, Egner T, Greene M, Koechlin E, Mangels J, Hirsch J (2006) Predictive codes for forthcoming perception in the frontal cortex. Science, 314(5803):1311-1314.

47. Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit Perceptual Anticipation Triggered by Statistical Learning. J. Neurosci., 30(33), 11177-11187.

48. Rugg MD, Johnson JD, Park H, Uncapher, MR (2008) Encoding-retrieval overlap in human episodic memory: A functional neuroimaging perspective. Progress in Brain Research 169: 339-352.

49. Tubridy S, Davachi L (2011) Medial Temporal Lobe Contributions to Episodic Sequence Encoding. Cereb. Cortex 21(2): 272-280.

50. Davachi L (2006) Item, context and relational episodic encoding in humans. Current Opinion in Neurobiology 16(6): 693-700.

51. Eichenbaum H (2000) A cortical-hippocampal system for declarative memory. Nature Rev. Neuroscience 1(1): 41-50.

52. Devan BD, White NM (1999) Parallel information processing in the dorsal striatum: relation to hippocampal function J. Neurosci. 19(7): 2789–2798

53. Rosenbaum RS, Ziegler M, Winocur G, Grady CL, Moscovitch M (2004) I have often walked down this street before: fMRI studies on the hippocampus and other structures during mental navigation of an old environment. Hippocampus 14(7): 826-35.

54. Dusek JA, Eichenbaum H (1997) The hippocampus and memory for orderly stimulus relations. Proc Natl Acad Sci USA 94:7109–7114

55. Chun MM, Phelps EA (1999) Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. Nat. Neurosci. 2: 844 - 847

56. Frank MJ, Rudy JW, O'Reilly RC (2003) Transitivity, Flexibility, Conjunctive Representations and the Hippocampus: II. A Computational Analysis. Hippocampus 13: 341-354

57. Norman KA, O’Reilly RC (2003) Modeling Hippocamapl and Neocortical Contributions to Recognition Memory: A Complementary-Learning-Systems Approach. Psychological Review. 110(4): 611-646

58. Lisman J, Redish AD (2009) Prediction, sequences and the hippocampus. Phil. Trans. R. Soc. B. 364:1193-1201

59. Gluck MA, Myers CE (1993) Hippocampal Mediation of Stimulus Representation: A Computational Theory. Hippocampus 3(4): 491-516.

60. Harrison LM, Duggins A, Friston.KJ (2006) Encoding uncertainty in the hippocampus. Neural Networks 19: 535-546.

61. Schott BH, Sellner DB, Lauer C-J, Habib R. Frey JU, et al. 2004) Activation of Midbrain Structures by Associative Novelty and the Formation of Explicit Memory in Humans. Learn. Mem. 11: 383-387.

62. Hikosaka O, Bromberg-Martin E, Hong S, Matsumoto M (2008) New insights on the subcortical representation of reward. Current opinion in neurobiology18(2): 203-208.

63. Friston KJ, Danizeau J, Kiebel SJ (2009) Reinforcement Learning or Active Inference? PLoS ONE 4(7): e 6421. Available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0006421. accessed: 12/22/11.

64. Oldfield RC (1971). The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9: 97–113.

65. Lohmann G, Mueller K, Bosch V, Mentzel H, Hessler S, et al. (2001) Lipsia - a new software system for the evaluation of functional magnetic resonance images of the human brain. Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society 25 (6):449-457.

66. Talairach J, Tourneaux P (1988). Co-planar stereotaxic atlas of the human brain. Stuttgart: Thieme

67. Worsley KJ, Friston KJ (1995) Analysis of FMRI time series revisited - again. Neuroimage 2 (3):173-81.

68. Nichols TE & Holmes AP (2001) Nonparameric Permutation Tests For Functional Neuroimaging: A Primer with Examples. Human Brain Mapping 15:1-25.

## Figure Legends

Figure 1: Reconstructed, color-coded caudate ROIs in a 3-D rendered brain.

Figure 2: Reconstructed, color-coded hippocampal ROIs in a 3-D rendered brain.

Figure 3: Reconstructed, color-coded habenular (A) and nigral (B) ROIs in a 3-D rendered brain.

Figure 4: Modelled BOLD for surpise over the iterations of unpredictables in the fMRI session. I3: surprise for the 3 times pre-exposed; I6: surprise for the 6 times pre-exposed; I9: surprise for the 9 times pre-exposed unpredictables.

Figure 5: Modelled BOLD for entropy over the iterations of unpredictables in the fMRI session. H3: entropy for the 3 times pre-exposed; H6: entropy for the 6 times pre-exposed; H9: entropy for the 9 times pre-exposed unpredictables.

## Figure Legends for Review (supplementary material)

Figure S1: Display of the hippocampal ROIs taken from a non-rendered two-dimensional brain image, at Talairach coordinate y = -29.

Figure S2: Display of the left hippocampal ROI taken from a non-rendered two-dimensional brain image, at Talairach coordinate x = -28.

Figure S3: Display of the caudate ROIs taken from a non-rendered two-dimensional brain image, at Talairach coordinate y = 9.

Figure S4: Display of the right caudate ROI taken from a non-rendered two-dimensional brain image, at Talairach coordinate x = 12.

Figure S5: Display of the habenula ROIs taken from a non-rendered two-dimensional brain image, at Talairach coordinate z = 3.

Figure S6: Display of the substantia nigra ROIs taken from a non-rendered two-dimensional brain image, at Talairach coordinate y = -20.

## Tables

Table 1: Overview of conditions and exposition numbers

| Condition | No. of different scripts of the condition | Preexposition number | Iterations in fMRI session | Repetitions of original (pre-fMRI version) during fMRI | Repetitions of identical version within fMRI session |
|---|---|---|---|---|---|
| *Originals* | 9 | 3, 6, or 9 | 9 | 9 | 9 |

| | | | | | |
|---|---|---|---|---|---|
| *Divergents* | 9 | 3, 6, or 9 | 9 | - | 9 |
| *Unpredictables* | 9 | 3, 6, or 9 | 5 | - | 1 |
| *New Originals* | 3 | - | 9 | - | 9 |
| *Singletons* | 6 | - | 1 | - | 1 |

Figure 1
Click here to download high resolution image

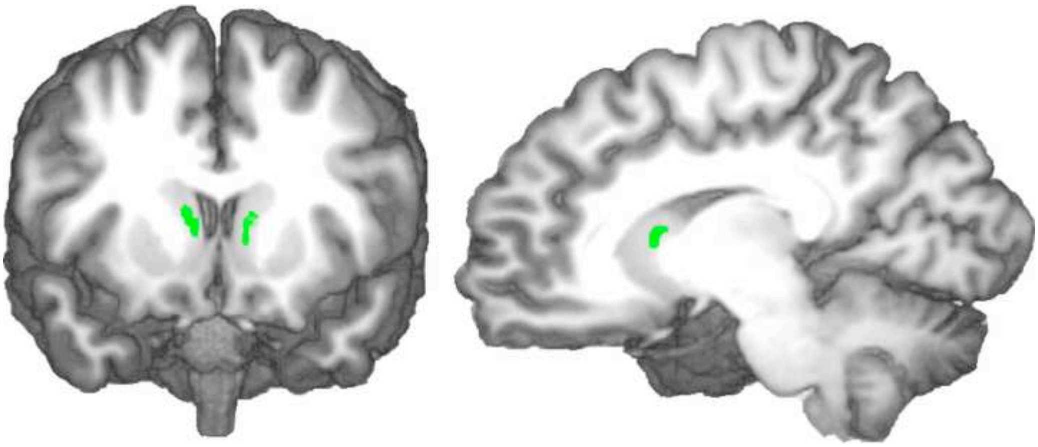


Figure2
Click here to download high resolution image

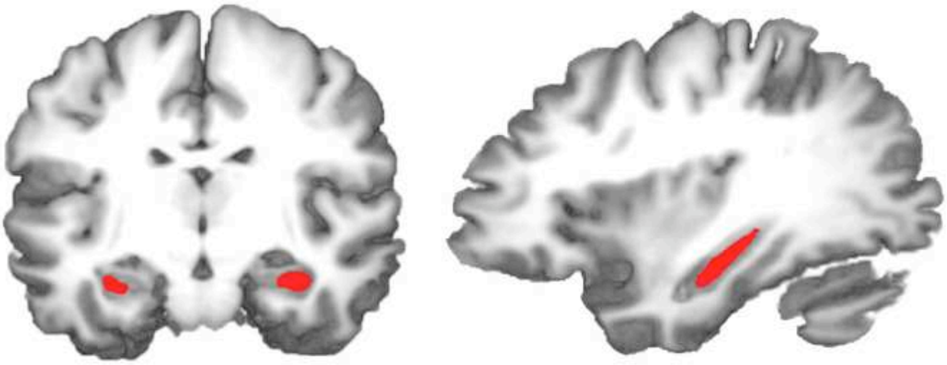


Figure 3

Click here to download high resolution image

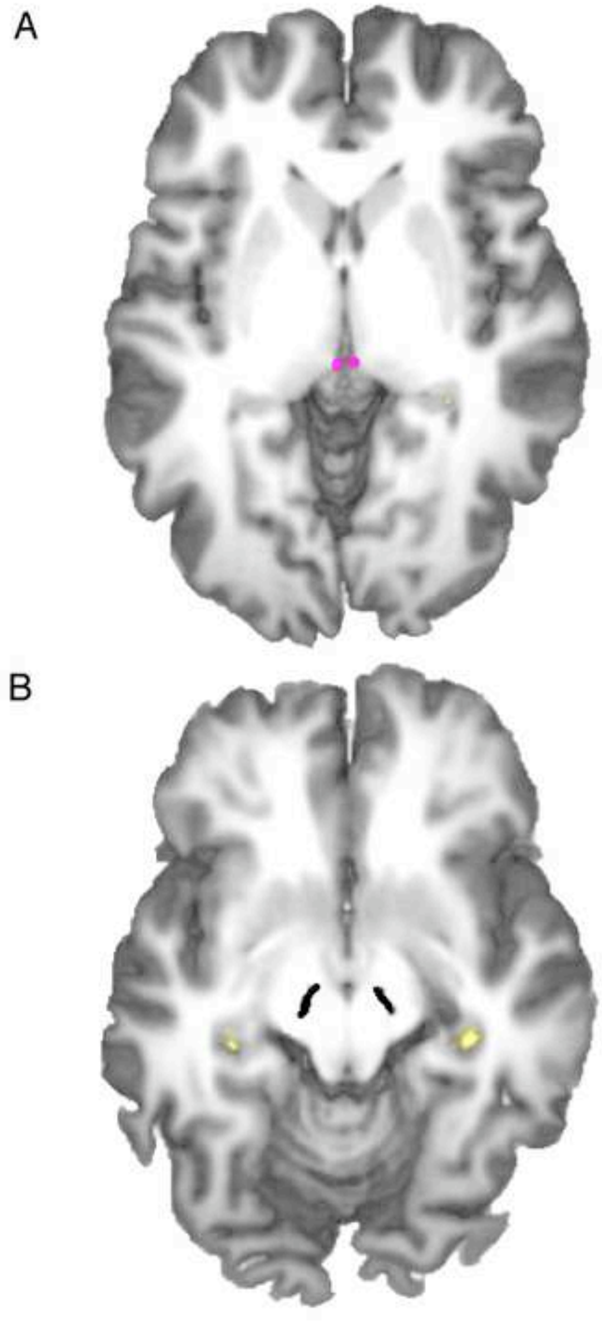


Figure 4
Click here to download high resolution image

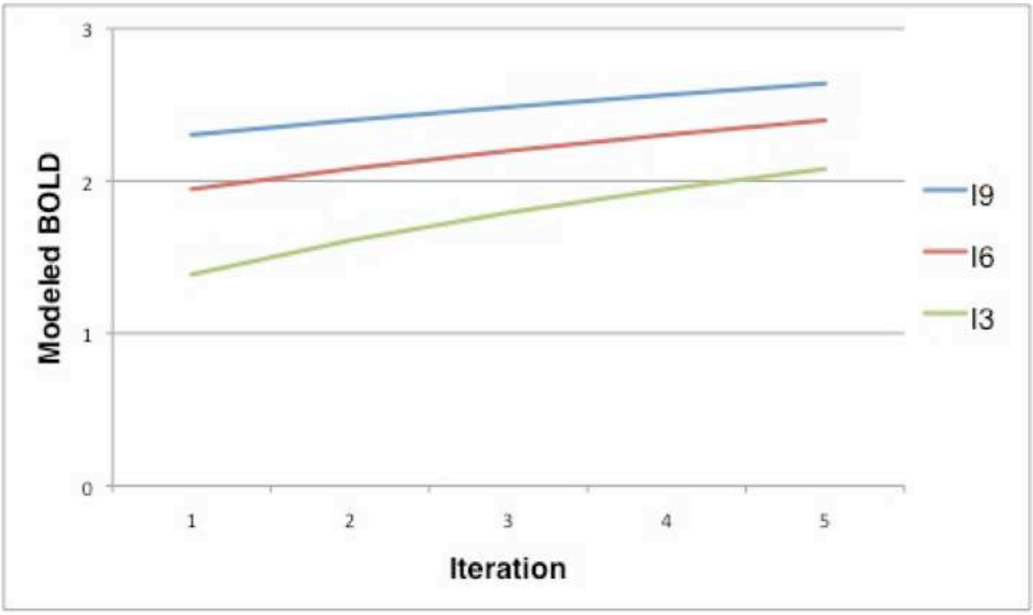
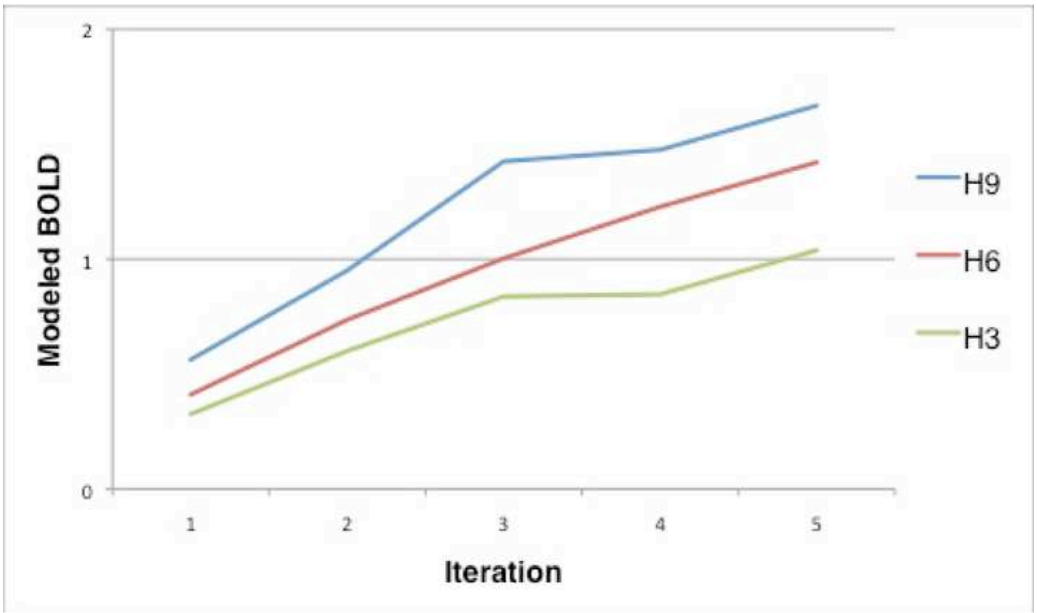
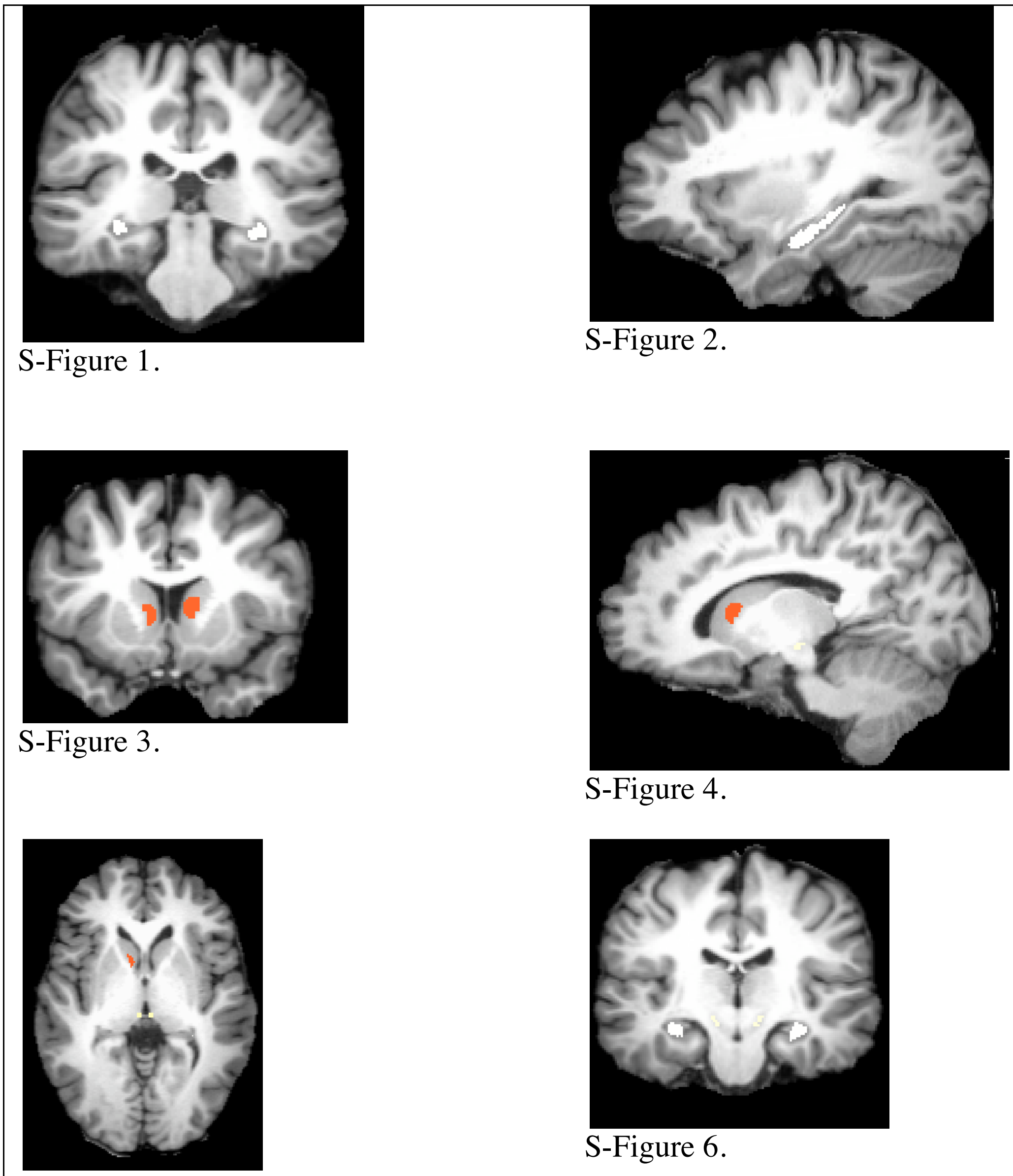


Figure 5
Click here to download high resolution image





S-Figure 1.

S-Figure 2.

S-Figure 3.

S-Figure 4.

S-Figure 5.

S-Figure 6.

Figure 1: The encompassed figures were added as Supplementary Material to the submission to allow an unbiased evaluation of the handdrawn ROIs. They are presented separately in this graph as they do not appear in the ‚Authors' proof' that is encompassed above. S-Figure 1: Hippocampus y = -29; S-Figure 2: Hippocampus x = -28; S-Figure 3: Caudate y = 9; S-Figure 4: Caudate x= 12; S-Figure 5: Habenula z = 3; S-Figure 6: Substantia nigra y = -20.

# 3 Discussion

## 3.1 Prediction Errors in the Basal Ganglia

The main aim of the first described study, the “Caudate study” (Schiffer & Schubotz, 2011) was to investigate whether the neural correlates of prediction errors that are unrelated to reward could be established in the striatum. The history of the establishment of prediction errors in the basal ganglia in animal studies (Ljungberg, Apicella, & Schultz, 1992; Montague, et al., 1996; Schultz, 2000; Schultz et al., 1997) seems to have imposed a reward-related view on future studies (O’Doherty et al., 2006; O’Doherty, Dayan, Friston, Critchley, & Dolan, 2003; O’Doherty et al., 2004). In fact, it was Watson (Watson, 1913) himself who stated: “The man and animal should be placed as nearly as possible under the same experimental conditions. Instead of feeding or punishing the human subject, we should ask him to respond by setting a second apparatus until standard and control offer no basis for a differential response.” The basal ganglia are in fact supposed to have access to representations of context (Saint-Cyr, 2003; Goldberg, 1985), bind information from different cortical areas (Graybiel, 1998), dispose of neurons with predictive capacities (Aosaki et al., 1994), and are supposed to receive efference copies from motor commands (Alexander et al., 1995). Thus there is ample evidence suggesting to leave the reward-centred view and investigate if the basal ganglia may be involved in not-reward related, perceptual prediction errors, especially if concerned with action observation.

With two notable exceptions (den Ouden et al., 2010; den Ouden, Friston, Daw, McIntosh, & Stephan, 2009), the presented studies are the first to show perceptual prediction errors in the striatum (Schiffer & Schubotz, 2011; Schiffer, Ahlheim, Wurm, & Schubotz, submitted). In the following I will describe two main issues that relate to the results of the conducted studies: First, what are the prerequisites to learning from prediction errors? And second, what determines the learning rate and the content of learning from prediction errors? The first part of the discussion will focus on the former issue and deal with internal forward models in the basal ganglia. The latter issue will be covered in the second main chapter of the discussion centred on actual adaptation of internal forward models. The experiments will be referred to as “Caudate”(Schiffer & Schubotz, 2011), “Bias” (Schiffer, Ahlheim, Ulrichs, & Schubotz, in press) and “Entropy” experiments (Schiffer, Ahlheim, Wurm, & Schubotz, submitted). Since the results discussed in the Bias and Entropy articles derive from the same experiment, I will use the singular (‘study’) to refer to the underlying experiment. I will outline within each discussion how empirical studies could clarify research questions that evolve from my results and the backdrop provided by relevant literature.

## 3.2 Internal Forward Model Projections in the Basal Ganglia

If the basal ganglia are capable of coding for prediction errors, they must have some sort of access to current internal forward models. Indeed, the overlap in the choice of terms of motor control theory (Wolpert & Kawato, 1998; Wolpert & Miall, 1996) that posits that internal models are generated to predict the next internal state achieved by a motor command, and that of modern models of basal ganglia function is striking (Bischoff-Grethe, et al., 2002; Redgrave et al., 1999). Already Alexander and colleagues proposed that it is an efference copy which travels through the cortico-basal ganglia-thalamo-cortical loops (Alexander et al., 1986). Redgrave and colleagues proposed that the striatum receives copies of the commands sent to the motor plant (Redgrave et al., 1999), another term from motor control theory (Wolpert & Miall, 1996). Lastly, Bischoff-Grethe and colleagues described a theory according to which the direct pathway in the cortico-basal ganglia-thalamo-cortical loops computes predictions of the next sensory state (Bischoff-Grethe et al., 2002). Hence, the idea that internal forward models are generated in the cortico-basal ganglia-thalamo-cortical loops deserves clarification.

### *3.2.1 Neuroanatomical Considerations Concerning Internal Forward Models*

The direct pathway does not seem to have direct access to the spinal cord (hence the historical name “extrapyramidal motor system” (cf. Mink, 1996). Thus, a traversal of a motor command through the direct basal ganglia pathway seems at least inefficient. This is one reason why it has been proposed that it is an efference copy, I call it internal forward model, which is generated through the cortico-basal ganglia-thalamo-cortical loops (Alexander et al., 1986; Bischoff-Grethe et al., 2002; Redgrave et al., 1999). The indirect pathway has output to the spinal cord via brain stem nuclei (Takakusaki, Saitoh, Harada, & Kashiwayanagi, 2004; cf. Bischoff-Grethe et al., 2002), and, as laid out above, the indirect pathway also back-projects to the cortex (Alexander et al., 1986; Haber, 2003; Parent & Hazrati, 1995a/b; Smith et al., 1998). I therefore propose that internal forward models, for example of the next sensory state of alternative models, are generated via the indirect pathway. Graybiel described predictive properties of large medium spiny neurons in the striatum, neurons that are related to the direct as well as the indirect pathway (Aosaki et al., 1994; Graybiel, 1998). Graybiel and colleagues have discussed the possibility that the striatal matrix compartments act as templates that learn to associate input from different cortical patterns (cf. ‘chunking’). This could deliver a powerful mechanism for the establishment of internal forward models, which rely on the association of a motor command, an action, or a choice - depending on the level of abstraction - with a sensory state. On the motor level, for example, the projections of the handknob in area M1 interdigitate in the striatum with the projections from the hand area in S1 (Graybiel, 1998; Mink, 1996). “Generation” and weighting of a forward model would therefore rely on a binding of the cortical activations, and thus association, of either area within a defined time window (Morris, Arkadir, Nevet, Vaadia, & Bergman, 2004). The activity within one cortical area would according to this theory activate the striatal compartment which the cortical area projects to, where it would lead to the activation of the associated input from different areas. We thus find that the basal ganglia have the neuroanatomical setup to generate internal forward models. If this mechanism was trained via dopaminergic processes as I will describe shortly, it would result in associations of different strengths, i.e., weighted internal forward models.

### *3.2.2 Decision Making Theory and Internal Forward Models in the Striatum*

As elaborated in the chapter *Basal Ganglia Pathways (1.8.2)*, research on decision making or action selection in the basal ganglia has shown that dopaminergic innervation of the striatum fosters activity in the direct pathway and suppresses activity in the indirect pathway. Redgrave and colleagues have recast the basal ganglia as “a solution to the selection problem”, as the fostering of activity in the direct pathway could lead to the response in favour of the representation in the direct pathway. But how does decision making or selection relate to prediction? Ideomotor control theory posits that actions are chosen based on associated consequences (Herwig & Waszak, 2009; Herwig, Prinz, & Waszak, 2007; James, 1890; Krieghoff, Waszak, Prinz, & Brass, 2011; Kühn, Seurinck, Fias, & Waszak, 2010; Waszak et al., 2005). If we understand consequences as sensory states that are achieved through actions, this unifies the action selection account with the prediction of anticipated sensory states. Lets regard an example of activity in the primary motor cortex. Loosely speaking, the account of anticipated sensory states pertains to saying that the projections through the cortico - basal ganglia-thalamo-cortical loops do not carry a copy of the activity, that is associated with a movement (‘motor command’), but lead to activation of the representation of the outcome of the motor command as can be registered in sensorimotor areas. This description in essence refers to a projection that allows prediction of sensorimotor consequences. (Of course, this relates to the motor loop, more cognitive percepts could be located in different loops).

If internal models are projected via the cortico-basal ganglia-thalamo-cortical loops, this opens up the opportunity for weighted forward models. Weighted forward models are discussed in decision making theory but also in the predictive coding account. Each time a model, in the case of decisions the representation of a choice, is performed, LTP leads to a fostering of this model. At the same time, all concurrently present alternative models are weakened through the D2 driven LTD mechanisms in the indirect pathway. This stamping in of response patterns ultimately pertains to weighted forward models (Frank, 2006; Graybiel, 1998).

Thus, we find that associative cortices’ projections must play a substantial part in shaping and activating internal forward models. But cortical input structures concerned with the instantiation of motor activity are relevant, too.

### *3.2.3 The Supplementary Motor Area: Internally Triggered Forward Models*

As I have described in the review of cortico-basal ganglia-thalamo-cortical loops, the supplementary motor area (SMA) is considered an important input structure to the basal ganglia. The SMA sends bilateral projections to the striatum and has in fact been associated with predictive, intentional action selection (Goldberg, 1985). Its neural coding seems to concern precompiled action representations (Goldberg, 1985). Precompiled action representations rely on access to representations of the context of an action and to an internal model of the desired outcome “for internal error correction” (Goldberg, 1985). Of particular interest to the idea that internally initiated forward modelling involves the basal ganglia is the finding that repeated internal emulation of a movement can result in an improved performance of the movement (Jeannerod, 1995; Yágüez, Canavan, Lange, & Hömberg, 1999; Yágüez et al., 1998) even after stroke (Liu, Chan, Lee, & Hui-Chan, 2004) and in Huntington’s disease (HD) patients (Yágüez et al., 1999). At the same time, skilled, near automatic performance involves the basal ganglia (Floyer-Lea & Matthews, 2004). FMRI could be employed to test whether mental imagery training results in a similar pattern of activity change as suggested for motor training, i.e., a progressive involvement of the basal ganglia (Floyer-Lea & Matthews, 2004; Jueptner & Weiller, 1998).

One study that did investigate the comparison between mental imagery training and motor training did not report striatal activity (Nyberg, Eriksson, Larsson, & Marklund, 2006). However, the design of this particular study may have been suboptimal, as the mental imagery condition demanded the participants to cross their fingers and look at them while imagining finger tapping. This kind of proprioceptive and visual feedback is known to subdue imagery effects. Another aspect is that the fMRI study measured activity pre and post training, but not during the process itself. Training related basal ganglia activity is known to decrease (Juepner & Weiller, 1997), possibly accompanied by a ‘hard-wiring’ of the motor programme in the cerebellum (Hikosaka, Nakamura, Sakai, & Nakahara, 2002). In fact, Nyberg and colleagues did report cerebellar activity increase over the course of training by imagery (Nyberg, et al., 2006). Lastly, Doya proposed that basal ganglia learning is error and reward driven, while cerebellar learning is not (Doya, 1999). The influence of errors and rewards on the learning process in this study cannot be determined (Nyberg, et al., 2006). In sum, the suggestions made in the seminal Goldberg article (1985), taken together with these clinical results, indicate that internal forward models that are internally triggered, possibly in the SMA are generated in the basal ganglia.

### *3.2.4 Lateral Premotor Cortex Predictions*

While the SMA is associated with internally guided prediction, the lateral premotor cortex (PM) is associated with the prediction of external events and external actions (Schubotz, 2007). Decety and colleagues (Decety et al., 1997) proposed that PM activity is associated with action observation if no imitation is necessary, whereas action imitation and observation for imitation were to draw on the SMA (Decety et al., 1997). Whether internally triggered or not, the predictions of the PM may also involve the basal ganglia. The cortical focus in action observation and imitation paradigms seems to have led to a neglect of basal ganglia contributions. Basal ganglia activity is often reported, but mostly without further comment (Aziz-Zadeh, Koski, Zaidel, Mazziotta, & Iacoboni, 2006; Buccino et al., 2004; Cross, Kraemer, Hamilton, Kelley, & Grafton, 2009; Iseki, Hanakawa, Shinozaki, Nankaku, & Fukuyama, 2008; Munzert, Zentgraf, Stark, & Vaitl, 2008; Ramsey & Hamilton, 2010). In fact, basal ganglia activity has even been shown to accompany the non-motor predictions of the premotor cortex (Bubic, von Cramon, Jacobsen, Schroger, & Schubotz, 2009; Schubotz & von Cramon, 2004). It is unclear, however, in how far the basal ganglia involvement relies on internally triggered predictions. The lack of research in this area is altogether surprising: the motor loop is the most investigated cortico-basal ganglia-thalamo-cortical circuit. Action observation and imitation are both supposed to rely on internal action models

(Jeannerod, 1995). In short, regarding the PM activity during prediction of external events (and actions), and basal ganglia activity in action observation paradigms, basal ganglia contributions could actually be relevant for both, internally triggered and externally triggered predictions and the issue deserves further empirical clarification.

### *3.2.5 Clinical Studies*

One study has reported that Parkinson's disease (PD) patients show little benefit from mental imagery (Yágüez et al., 1999), delivering an argument in favour of the contributions of the basal ganglia to the benefits derived from mental imagery. In contrast to PD patients, no such deficit was reported for Huntington's Disease (HD) patients (Yágüez et al., 1999). In this context, it is very interesting to note that the SMA projects at least evenly to putamen and caudate (Haber, 2003; Parent & Hazrati, 1995a), if not even preferentially to the putamen (Alexander & Crutcher, 1990; Alexander et al., 1986; Di Martino et al., 2008). It has been suggested that degeneration of the striatum in HD patients is at early stages more pronounced in the caudate (Sturrock & Leavitt, 2010; VonSattel et al., 1985) than in the putamen. Thus, it is at least possible that motor imagery benefits in HD are spared due to the SMA-putamen projections. The pattern of striatal degeneration, in this case in terms of dopamine depletion, in PD may be reversed, with the putamen being earlier affected than the caudate (Kish, Shannak, & Hornykiewicz, 1988). The hallmark of PD is dopamine depletion in the substantia nigra. This depletion could affect long-term potentiation of target models as well as long-term depression of non-target models (Frank, 2006) in the two pathways. This delivers a powerful account of impairment of learning from motor imagery in PD: the relevant internal forward model of the action would simply not be sufficiently fostered by LTP. On a related note, additional support for the hypothesis that the cortico-basal ganglia-thalamo-cortical pathways are involved in the internal modelling of succeeding states comes from the finding that Parkinson's disease patients do not show predictive strategies in motor tasks (Crawford, Goodrich, Henderson, & Kennard, 1989; Flowers, 1978).

With regard to action observation, and thus externally triggered, putatively PM dependent predictions, it has been shown that PD patients do not suffer a specific deficit in moving their arms incongruently to an arm movement displayed on a video screen compared to healthy controls (Albert, Peiris, Cohen, Miall, & Praamstra, 2010). The authors conclude that this finding indicates that the "mirror neuron system" remains intact in PD. Even without the notion of a mirror neuron system, this finding supports a suggestion from the seminal 1985 Goldberg review (Goldberg, 1985). Goldberg proposed that PD patients can rely on involvement of their lateral premotor system to compensate for the compromised medial premotor system that he proposed to rely gradually more on the SMA and basal ganglia. Substantial support for the assumption that the internal models in the motor circuit, but possibly also other cortico-basal gangli-thalamo-cortical circuits, do not only carry internal forward models of actions, but also forward models according to the premotor cortex predictions of external (even non-biological) events comes from neuroimaging (Schubotz, 2007; Schubotz & von Cramon, 2003, 2004) and clinical studies (Schubotz & Sakreida, unpublished results). These studies showed an involvement of the basal ganglia in premotor predictions (Schubotz & von Cramon, 2004), basal ganglia activity towards violations of these predictions (Bubic et al., 2009; Bubic, unpublished results), and an impairment of these predictions in a clinical population with premotor and basal ganglia strokes (Schubotz & Sakreida, unpublished results). A recent clinical study suggested that 'the mirror neuron systems is mirrored in the basal ganglia" (Alegre et al., 2010). A more parsimonious, and very well testable account would be that premotor predictions (Bubic et al., 2009; Schubotz, 2007; Schubotz & von Cramon, 2003; Schubotz & von Cramon, 2004), including premotor predictions of external actions (Schubotz & von Cramon, 2004) are generated via cortico-basal ganglia-thalamo-cortical loops.

### *3.2.6 A proposal of a Clinical Study to Test Implications*

An internal forward model account of the cortico-basal ganglia-thalamo-cortical loops could be tested by a simple clinical study involving only behavioural measures and implementing an action observation paradigm. If the *initiation* of an internal model relies on the cortico-basal ganglia-thalamo-cortical loops, PD patients should be impaired in generating such a model to predict the course of an action. As a first step, PD patients could be exposed to action models in a similar fashion to the pre-exposition task I used in the experiment described in the “Bias” and “Entropy” articles. They could afterwards be confronted with a forced choice task. Within this task they would watch the beginning of the known action movies, which would suddenly stop, followed by the presentation of two photographs of possible endings of the movie. The PD patients would have to choose which photograph corresponds to the known action model. If they were delayed in this choice when depleted of dopaminergic medication, compared to a medicated state and to healthy controls, this would indicate that the generation of the predictive internal model relies on a basal ganglia projection. Additionally, one could test for their ability to detect prediction errors, when confronting them with the altered movie versions (compare the “Bias”-experiment) in a signal detection task that demands responses towards surprising movie developments. This would then figure as a test of compromised prediction error signalling, putatively dependent on intact basal ganglia functioning, in PD.

On a related note, it has been proposed that action segmentation relies on perceptual prediction errors and would be compromised in PD. However, so far no evidence for the involvement of the basal ganglia in action segmentation has been established (cf. Schubotz, Korb, Schiffer, Stadler, & von Cramon, submitted; Zacks & Swallow, 2007; Zacks et al., 2001). But when participants had to predict what would happen five seconds after an event boundary, the midbrain dopaminergic system has been shown to be selectively activated (Zacks, Kurby, Eisenberg, & Haroutunian, 2011). This finding is in line with an account of enhanced basal ganglia activity when the forward models are internally triggered. It would therefore be interesting to implement action segmentation in the initial pre-exposition phase of the proposed experiment, to test for performance in non-medicated PD compared to the medicated state and healthy controls.

## 3.3 Midway summary

To sum up shortly, it seems likely that one function of the cortico-basal ganglia-thalamo-cortical loops is to generate weighted internal forward models of upcoming sensory or motor states. The powerful dopaminergic learning mechanisms, i.e., LTP and LTD, can explain how tight associations between different inputs from a number of cortical areas can be formed. These associations can bind for example motor commands to resultant sensory inputs, thus building internal forward models. Importantly, the bivalent projections of the direct and indirect pathway enable a probabilistic weighting of competitive alternative models. In relation to the predictive coding account of action observation, use of these weighted models could pose a potential way to derive at the most valid predictions (in the temporal domain) for the currently perceived action. Parenthetically, the account given, stresses the importance of the basal ganglia loops for learning of internal forward models and choice of internal forward models. But this must not be taken to diminish the role of prefrontal cortex in the inhibition of actions, or in the overruling of prepotent but incorrect responses under conflict. In fact, the hyperdirect pathway that offers direct influence of the mesial prefrontal cortex to the subthalamic nucleus (Frank, et al., 2007) could serve just that: Prefrontal modulation of basal ganglia selection that overcomes the learned (heavily weighted) forward model.

## 3.4 Learning from (Prediction) Errors

The term prediction error as I use it describes a neural response to perceptions that deviate from expectations. Errors, however, are usually understood in a broader, more valence-centred fashion. It is a longstanding topic of psychological research that if humans commit errors, they intend to correct them, or at least try not to repeat them in the future (Thorndike, 1927).

Prediction errors, as described in the predictive coding account, are not necessarily understood to relate to the valence of an experience (but see Friston, et al., 2009). However, they are understood to change the underlying models' predictions, which will lead to less prediction errors after learning (Friston, 2002). The "Bias"-study dealt with the question how prediction errors influence learning. As hypothesized, adaptation rate depends on the previous solidity of a model. However, this study also opened up new avenues, as it suggested that especially states of relative certainty in a model allow differential processing of incoming information. This finding is highly interesting, when considering that the state we defined as 'bias' is anti-proportional to entropy. Balanced states, i.e., states in which the evidence for each model is equal, are of high entropy. Biased states, which dispose of a solid and a weak model, are signified by small Shannon entropy. The "Bias" and "Entropy" study indicated that different brain areas code for these states of high or low entropy. The experiment also quite successfully related the hippocampal response pattern towards entropy to a decrease in free-recall rate. Thus, we found that learning from prediction errors takes place, and that it is influenced by the reliability of the predictions-violating input. The fact that the anterior cingulate cortex (ACC) was activated for states of high bias, or low entropy, is incidentally also an uncommented finding by Harrison and colleagues (Harrison, Duggins, & Friston, 2006). This relates in an important fashion to different concepts of ACC function. The ACC has long been associated with monitoring for error commission or monitoring for conflict (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Botvinick, Cohen, & Carter, 2004; Holroyd et al., 2004; Holroyd, Yeung, Coles, & Cohen, 2005; Jocham & Ullsperger, 2009; Ullsperger & von Cramon, 2003, 2004, 2006; Yeung, Botvinick, & Cohen, 2004). I will argue that our finding from a paradigm of perceptual prediction errors is relevant to the latter proposed function and expands it substantially. Lastly, this discussion proposes that the relationship between prediction errors and errors that an individual commits (or witnesses, as we will see), needs further investigation.

### *3.4.1 The Relationship between Errors and Prediction Errors*

Corrections of the errors an individual commits may be accompanied or driven by the feeling of annoyance at the error (Thorndike, 1927). To investigate the error-characteristic of a prediction error, while interesting in itself, is further suggested by the presence of habenular activity in both presented experiments ("Caudate" and "Entropy" experiment). Activity in the habenula is associated with punishment and outcomes that are worse than expected (Matsumoto & Hikosaka, 2008). If its activity level was shown to reliably accompany prediction errors, this could mean that failed predictions are of a negative valence – possibly driving corrective responses.

### 3.4.2 Error-related Research

Experimentally, a number of different methods have been used to investigate errors. Ways to induce error commission are speeded response tasks (Garavan, Ross, Murphy, Roche, & Stein, 2002; Holroyd, et al., 2005; Shane, Stevens, Harenski, & Kiehl, 2008), tasks that demand a particularly difficult decision e.g. on degraded stimuli (Hughes & Yeung, 2011; Summerfield, Egner, Mangels, & Hirsch, 2006b), or when conflicting information is present (Holroyd & Coles, 2002; Holroyd et al., 2005; Hughes & Yeung, 2011; Potts, Martin, Kamp, & Donchin, 2011; Ullsperger & von Cramon, 2006; Yeung, Botvinick, & Cohen, 2004). Another way to investigate error responses is arbitrary feedback, for example in probabilistic (Holroyd, Krigolson, Baker, Lee, & Gibson, 2009; Holroyd & Coles, 2002), or guessing and estimation tasks (Oliveira, McDonald, & Goodman, 2007; Ullsperger & von Cramon, 2003). This form of feedback can for example be negative, and thus imply an error, when the participants response was in fact correct (Oliveira et al., 2007).

### 3.4.3 The ACC in Error Research

Error responses under conflict have been shown to activate the ACC and elicit a specific type of event related potential (ERP), the error-related negativity (ERN; Gehring, Goss, Coles, Meyer, & Donchin, 1993). The ERN was originally interpreted as an internal error-detection correlate, which is not dependent on external feedback (Gehring et al., 1993, cf. Ullsperger & von Cramon, 2003). On the other hand, it has been suggested that the ERN and associated ACC activity reflect the fact that a response is made in the presence of conflicting information. The ERN is proposed to stem from the processing of the evidence for the alternative (correct) response, after the (incorrect) response has occurred (Yeung et al., 2004). Thus, it is still under debate whether the

ERN (and ACC activity) results from the internal detection of the commission of errors, or from the conflict that follows errors, when the representation of the correct response garners strength (Yeung, et al., 2004). There is strong evidence in favour of the second theory, which dictates that the ACC activity and corresponding ERN accrue when two representations compete and one wins dominance over the other (Yeung, et al., 2004, 2004). It is thus reasoned that if a premature incorrect response is made, and evidence for the correct response accumulates thereafter, the ERN is issued (Yeung, et al., 2004). One of the most convincing findings is that correct responses are preceded by an ERP, the N2, which is similar to the ERN, but occurs before the response (Yeung, et al., 2004). Thus, it seems that if the representation of the correct response surpasses that of the conflicting response, the N2 is elicited and a correct response is made. Evidence in favour of the theory is further delivered by the fact that masking of the conflicting stimulus material inhibits the ERN. This is putatively the case, because further processing of the representations, amounting to a stronger representation of the (unchosen) correct response, is prevented (Hughes & Yeung, 2011). This account of ACC activity is very much in line with the proposed interpretation in the “Bias” article. I propose to use the term bias instead of conflict as bias seems more suited to describe unbalanced states in favour of the dominant response, such as in the described experiment as well as in the case of the N2 for correct responses. The description of two parallel representations of unequal strength is highly reminiscent of the account weighted internal forward models of cortico-basal ganglia-thalamo-cortical loops I presented. The relationship between the ACC and the basal ganglia with regard to bias between forward models remains to be established.

Negative feedback, even if invalid, also elicits a negative ERP, which has been called f-ERN. In analogy to the widespread interpretation of the ERN as an error-correlate, it has been suggested that the f-ERN indicates that feedback suggests that an error has been committed (Gehring, et al., 1993). This response to invalid feedback, however, depends on the existence of valid predictions on what the feedback should be like (Holroyd, et al., 2009; Oliveira et al., 2004). These findings have stipulated the hypothesis that the f-ERN is in fact a response to outcomes that are different from what was expected (Oliveira et al., 2004). The idea that the f-ERN demands the existence of valid predictions (Holroyd et al., 2009) likewise relates to the existence of bias states. If no solid model of the correct response existed, i.e., no bias towards a response was present, no f-ERN is issued (cf. Holroyd et al., 2009; cf. Donkers, Nieuwenhuis, & van Boxtel, 2005). However, since the ERN and f-ERN have been proposed to differ slightly in scalp distribution (Potts, et al., 2011) it is possible that different subareas of the ACC concurrently code for both, biased states (reflected in EEG in the ERN) and perceptions that deviate from biased states (reflected in EEG in the f-ERN). A simple experiment that varies both, the anticipated reliability of feedback and the difficulty of a discrimination task could test this hypothesis. In an easy discrimination task, incorrect feedback should not lead to enhanced control processes in the next trial, while a difficult task should shift attention towards feedback even if not absolutely reliable.

### *3.4.4 Neuroanatomy and Neurotransmitters of Errors*

The ERN/f-ERN and accompanying ACC activity have been associated with the midbrain dopaminergic system (Holroyd & Coles, 2002; Ullsperger & von Cramon, 2006). Specifically, Holroyd & Coles (2002) suggested that a decrease in dopaminergic activity in the midbrain causes the ERN/f-ERN. This assumption fits with data from Ullsperger and von Cramon (2006), who could show, that the habenular complex is activated when negative feedback is (unexpectedly) delivered. Habenula stimulation results in a massive decrease of dopaminergic output in the midbrain. Therefore, it has been suggested that the habenular influence on the midbrain dopaminergic system can code for outcomes that are worse than expected (Hikosaka et al., 2008; Hong & Hikosaka, 2008; Matsumoto & Hikosaka, 2007; Wickens, 2008). With regard to the habenula activity in the presented studies, outcomes that elicited habenular responses may have been different from what was expected, but not necessarily worse. On the one hand, habenular activity in the "Caudate" study (Schiffer & Schubotz, 2011) may have indicated that participants experienced the observed error like an error they had committed themselves, due to the prolonged behavioural training in the paradigm. On the other hand, this explanation does not fit the results presented in the "Entropy" study. Here, we found habenular activity for repeatedly unpredictable events, apparently unrelated to valence. The result indicates that outcomes that constitute prediction errors, even if not related to committed, factual errors, excite the habenula, signalling the need to adapt predictions. This stands in relation to the dual pathway account: a dip in dopaminergic firing, incited by habenular activity could simply diminish LTP for the representation in the direct pathway, to prevent a weight-gain of the model. To test this hypothesis, it would be necessary to investigate whether negative events such as errors increase activity in the habenula more than unpredicted events.

Surprisingly, a similar test seems due for the field of research centred on the ERN/f-ERN. As mentioned above, research into the f-ERN has suggested that the f-ERN can be elicited by states that are different from what was expected (Oliveira, et al. 2004), and does not necessarily depend on outcomes that are worse than expected (Holroyd & Coles, 2002). Meanwhile, experiments have been conducted to investigate whether the observation of others' errors elicit the same activity as own erroneous responses (Shane et al., 2008; de Bruijn, de Lange, von Cramon, & Ullsperger, 2009; van Schie, Mars, Coles, & Bekkering, 2004). However, it seems unlikely that other people's errors are ever fully predicted. Hence, if the f-ERN was elicited by outcomes that diverge from expectations, this description would fit unexpected negative feedback as well as observed errors. Thus, the established f-ERN corresponding ACC activity in the studies that investigated observed errors so far does not necessarily support the interpretation of the ACC as an error-detecting structure. ACC activity may just as well stem from unpredicted perceptions that violate a solid model. It is therefore astonishing that no study has scrutinized the difference between unpredicted observed errors and predicted observed errors.

### *3.4.5 Proposed Study to Dissociate Errors from Prediction Errors*

To disentangle the concept of prediction errors and committed errors, I therefore propose a study that would employ a behavioural training that emphasized the need to perform actions correctly and thus create a negative valence for deviations from this performance. Secondly, it would employ a video pre-exposition as in the study described in the "Bias" and "Entropy" articles to create the expectation that specific actions will be conducted erroneously, while other actions would be associated with a correct conduction. In the fMRI, half of the previously erroneously conducted actions would reappear performed correctly, while half of the previously correctly performed actions would reappear erroneously. First of all, and maybe most importantly given the apparent confound in the literature, the proposed study would allow dissociating the observation of predicted errors from unpredicted errors. Thus, it could be tested whether the neural activation to observed errors mirrors the neural activation of committed errors, due to existence of bias, both when committed errors are evaluated, as well as when surprising errors are observed. Secondly, surprisingly correct actions could be compared to surprisingly erroneous actions, investigating whether the negative valence associated with errors of commission (as opposed to prediction errors) drive the habenula, or whether the habenular response established in my studies implicates the structure’s involvement in general breaches of expectation.

On a last note concerning the proposed experiment, it could be possible to discern ACC function as prediction error related (as suggested by the f-ERN findings) or bias related (as proposed for the ERN). If ACC activity was responsive to conflict, we would expect activity in case of the predicted errors, unpredicted correct, and unpredicted error actions, but not for the predicted correct actions. If two models need to compete to elicit what I call the bias response, the behavioural training should make the correct action prepotent, while the pre-exposition can create a competing model. In case of the unpredicted errors, this competing model would be considerably fragile, as it only comes into existence during the fMRI session. For the predicted correct actions, there would be no model competition, as only one action model (the correct) would guide expectations. A prediction error interpretation of ACC activity, in line with the proposed f-ERN interpretation of a mismatch detector in biased states, would predict ACC activity for the unpredicted correct and unpredicted error actions, but not for their predicted counterparts. Since the ERN and f-ERN have been proposed to differ slightly in scalp distribution, the proposed experiment, using spatially sensitive fMRI measures could even point towards a diverging localization for both functions with in the ACC.

## 3.5 Applying Computational Models

### *3.5.1 Interpretative Concerns*

We could show that high bias, which translates to low entropy, is marked by a distinctive pattern of brain activity. Moreover, as presented in the "Entropy" study, the hippocampus displayed a BOLD response that accorded to the information theoretic formula for entropy. One description ('bias') is built on deductive reasoning, the other employs a mathematic formula. There is a certain appeal to finding a formula to describe neural responses. The beauty of the modelling approach lies certainly within the generation of testable hypotheses. However, it seems very desirable to determine to test whether any function, such as e.g. encoding, or weighing underlies the respective modelled response. To make the critique quite clear: one explanation for repetition suppression is the predictive coding account of perception (Summerfield, et al., 2008). This interpretation relies on the fact that statistically more probable events show larger repetition suppression than improbable events (Summerfield et al., 2008; Turk-Browne et al., 2010). These statistical dependencies indicate that in the concerned studies, repetition suppression is not solely due to neural exhaustion. The fact that a representation is predicted due to recent activation of an associated representation decreases its activity on subsequent trials. Repetition suppression could be used to investigate for example the response of a certain brain area (e.g. in the parietal cortex) to certain grip types. If a repetition of certain grip types leads to less activity than the presentation of a new grip, while other variables are accounted for, one could deduce that the area codes for grip types. In contrast, the deduction that this brain area is responsive to predictions from a higher-level area is comparatively uninformative: this assumption provided the basis of the experimental operationalisation and should hence not be the conclusion. The same rationale must apply to all modelling approaches that are used to predict neuronal responses. In so far as a study investigates only whether any model or algorithm accounts for BOLD response, there is the possibility that the region is more than an indicator of a state according to the algorithm. The interpretation of the exact function, which is subject to responses accounted for by the algorithm, must rely on abductive reasoning, if it was not specifically tested.

The “Entropy” experiment was the third study (cf. Harrison, et al., 2006; Strange, et al., 2006; Friston, 2005) in recent years to locate entropy responses in the hippocampus; so far no account has been put forward whether the hippocampus simply detects entropy, or whether any function of hippocampal firing can be related to entropy as well. We related the finding of entropy-responses in the hippocampus to the free-recall rates of the respective movies: Movies that developed unpredictably were not as often recalled as movies that reliably diverged from the original in the same way. Entropy increased monotonously for the unpredictably ending movies. But bias high bias means low entropy. As discussed in the bias article the reliably diverging movies display biased states at the beginning and end of learning. Bias is correlated negatively with low entropy. Therefore, these movies display states of low entropy – and better free recall rates than the unpredictably ending movies that display increasing entropy.

The same critique applies to the caudate response. While it may be possible that the response of the caudate nucleus relates to the unpredictability, or surprise of an outcome, it remains to be established whether this response in fact determines for example the revision of the internal models. An interesting result in that regard comes from den Ouden and coworkers (den Ouden et al., 2010), who could show that the correlate of surprise in the striatum determines cortical coupling between fusiform face area, parahippocampal place area and the premotor cortex. The authors argue that this result could mean that bottom up visual information influences premotor activity to a larger degree following surprising than predicted events. Surprise for a single item occurrence is higher in states of low entropy (while the accumulation of surprise signifies high entropy, rendering unpredicted outcomes less surprising). In contrast to the visuo-motor influence account put forward by den Ouden and colleagues (2010), if large surprise co-occurs with more cortical coupling, this could indicate that internal models in the basal ganglia are revised to a larger degree in states of low entropy. This alternative explanation is thinkable, as the den Ouden (2010) study used DCM to show that the putamen influences cortical coupling between the FFA, PPA and PM, but DCM does not provide a means to establish which cortical fibres are involved in the coupling. To test the hypothesis that surprising events actually change putative internal forward models in the basal ganglia, and do more so for more surprising events, it would be necessary to deliver evidence that learning rate varies with cortical coupling and adaptation correlate with the amount of surprise.

To summarize, while it is an interesting finding that neural responses can be described by information theoretic terms, the pitfall of declaring them indicators of the algorithm that is used to model the BOLD should be avoided. Just as well, the area's function, for example encoding, or fostering the influence of an internal model, may be subject to the mathematically described state. This distinction between being an indicator of a state that accords to an algorithm or encompassing a function that is modulated by states that can be described by the algorithm amounts to the difference between the questions: Is the structure responsive to a state vs. what does the structure do/ code for in a certain state.

### 3.5.2 Model Competition

The second concern about modelling approaches that needs to be addressed is the competition between models. As Harrison and colleagues put it, the fact that one model accounts for a neuronal response does not mean that it is the best model (Harrison et al., 2006). On the other hand, if a model fails to predict neuronal responses, there is always the danger of accepting the null hypothesis as a reflection of the non-existence of the effect in the population: another scientific pitfall (Cohen, 1994). This problem is quite evident in applications of the TD-algorithm (Montague et al., 1996; Pan, Schmidt, Wickens, & Hyland, 2005; Schmidt, 2005; Sutton & Barto, 1990). The TD-algorithm has very influentially contributed to the understanding of the midbrain dopaminergic system (Montague, et al., 1996; Schultz, 2007; Schultz, et al., 1997; Schultz & Dickinson, 2000). However, even within the same theoretical framework, different parameter values lead to different predictions (Montague et al., 1996; Pan, et al., 2005; Schmidt, 2005; Sutton & Barto, 1990). And while some attempts have been made to infer which parameter describes the neural correlates best (Pan, et al., 2005), comparative studies are still lacking. Importantly, different environments, as imposed by different experimental constraints, may yield that the parameters of the model, for instance its learning rate, could be dependent on external influences (Rushworth & Behrens, 2008). Quite a void is apparent concerning the application of the model to reward-free paradigms. This may be due to the bias towards reward paradigms rooted in the breakthrough of the model in animal research (cf. Schultz, et al., 1997). It is dissatisfactionary, nevertheless. As I have described in the Introduction, the TD-algorithm is capable of learning states. In fact, the sum of cortical activity that is computed by the model (Montague, et al., 1996), does not necessarily relate to reward. Reward is just an input that excites the model to an excessive degree. In other terms, reward is the  “desired input”, and the TD-algorithm learns to predict the states that lead to reward (Montague, et al., 1996). The “Entropy” experiment presented evidence that BOLD responses in the striatum can be modelled as perceptual prediction errors, using the information theoretic formula for ‘surprise’. But it would be interesting to compare reward-related prediction errors and perceptual prediction errors within one and the same experimental design. Thus, the rationale of competitive modelling applies to the comparison of different computational models, such as predictive coding and the TD-algorithm. In their 2009 study, den Ouden and colleagues applied a model based on the Rescorla-Wagner learning rule to striatal responses and found the model to reliably predict neural activity. In 2010, another paper by den Ouden and colleagues found surprise in terms of predictive coding (or Information Theory) to account for striatal responses. What seems due is a comparison of different models in their ability to predict for example striatal or basal ganglia responses; preferably, distinguishing between reward-related prediction errors and perceptual prediction errors.

## 3.6 In Psychological Terms

I have tried to evade using the term “sensation” in the presentation of predictive coding. In psychology, the distinction between perception and sensation has been discarded, after a long debate of dualism and the stage at which sensation is turned into perception, if the concepts are separable at all (Watson, 1913). Von Helmholtz in describing his principles of “Unbewusste Schluesse” (Helmholtz, 1866), i.e., unconscious conclusions (Helmholtz, 1866), maintained that sensations are the cause that leads the brain to infer perceptions. Not surprisingly, some authors that discuss predictive coding reliably mention the works of von Helmholtz (Friston, 2005). But predictive coding does not help to solve the debate, which is the main reason why I refrained from using the term sensation: Why is that?

If we follow the arguments of predictive coding, the recurrent projections between cortical hierarchies turn all activity patterns into "model-induced activity" rather than sensations, as they all are to a degree due to prediction. If predictions guide perceptions fallibly, it has been shown that visual areas according to the predictions, but unrelated to the external cause are activated (Summerfield, et al., 2006b). The term sensation seems difficult to accommodate, unless for the "lowest" level of the hierarchy. However, responses that seem to display characteristics of predictive coding have been established in the retina and the lateral geniculate nucleus (Huang & Rao, 2011; Friston, 2005). Moreover, if we accept the idea that spontaneous activity can code for the probability of a perception, there seems to be no room left for sensations, ie., unfiltered input. Every input relates to a degree to a generative model. This account suggests, that predictive coding should be evaluated with regard to whether newborns dispose of generative apriori models. But this is beyond the scope of the current thesis. To sum it up, predictive coding therefore does not elucidate at what point of neural coding the term sensation could be replaced by perception, hence I did not use the term sensation. But this is a matter of unfortunate terms. On a last note, other models of course discuss reasonable concepts of perception and sensation, to give an example related to predictive coding, linking perception to unimodal association cortices that provide categorical coding (Mesulam, 1998).

Regarding the problem of perceiving correctly or rather functionally correct, predictive coding proposes that perception yields the most probable, but not necessarily the correct representation of the cause (Friston, 2005).

I have mentioned in the Introduction that learning functional predictions enables to respond efficiently, but that not all perceptions are aimed at responses. The presented studies have engaged paradigms wherein prediction errors were not relevant to response decisions. It would be interesting to test whether rewarded or action–relevant learning from prediction errors leads to faster adaptation than what we have seen in the employed paradigms that contained no explicit reward or demand of response following prediction errors. Incidentally, this would relate to the predictions made by the TD-algorithm, proposing that reward is a "special" input to the system that leads to more weight-change.

## 3.7 Final Remarks

The presented experiments have delivered evidence that prediction errors of internal forward models of observed action are signalled in the basal ganglia. The results underpin the idea that weighted internal forward models are generated in the basal ganglia and can be used for perceptual inference, for example concerning observed actions. Responses to prediction errors, also in the form of model adjustment or learning, depend on the strength of the internal forward model and reliability of the incoming information.

# 4 Appendix

## 4.1 The Kullback-Leibler Divergence

The change of probabilities, ie. beliefs on the distribution of an event in the real world is usually captured in the Kullback-Leibler Divergence. This divergence describes the difference between the prior probability at observation n-1 and the posterior probability at n0. Remember that if an event is informative, ie. surprising, it changes the probability. Informative events lead to a larger Kullback-Leibler Divergence. Lets call the prior probability at n-1 Q(data) and the posterior probability at n0 P(data). The Kullback-Leibner Divergence is calculated as:

$$D(P;Q) = \sum P(data)\, log \frac{P(data)}{Q(data)}$$

(Doya & Ishii, 2007; Itti & Baldi, 2005; Kullback & Leibler, 1951). The Kullback-Leibler Divergence is sometimes referred to as surprise (Itti & Baldi, 2005). The use of the term for two concepts is unfortunate. While the information theory approach to surprise relates to the predictability of an observation, the Kullback-Leibler Divergence relates to the change in posterior probability initiated by surprise (Doya & Ishii, 2007; Itti & Baldi, 2005). The aim of perception has been described as the minimization of the Kullback-Leibler Divergence, since a minimal Kullback-Leibler Divergence indicates, that the prior belief needs no updating (van de Veen, & Schouten, 2001). If the Kullback-Leibler divergence approaches zero, events are not unpredictable anymore, indicating a sound internal model (in terms of belief) of the external world. At the outset of learning, the prior probability is unlikely to contain the correct beliefs on observations. Surprising observations occur, leading to high uncertainty (entropy) that becomes lower if the internal beliefs are revised according to the Kullback – Leibler Divergence.

## 4.2 The TD-algorithm

Each sensory state that occurs is represented in one state vector (x(i,t)). The length of the vector is a matter of definition. The length of the vector determines how many time steps the model can remember (Montague, et al., 1996). At a given time t, the component i of the vector is 1, if the event that activated the cortical domain i occurred t-1 time steps ago. All other components equal zero. If the event occurred more time steps in the past than the length of the state vector, all components are zero. This form of representation is called serial-compound stimuli representation (Montague, et al., 1996). For each state vector, there is a corresponding weight vector (w(i,t)). The weight vector has the same length as the state vector and represents the influence one state has on the current predictions. Usually, all weight vectors are set to zero at the beginning of learning (Montague, et al., 1996). The weight vector will come to determine how much influence a reward, or change in the sensory state has on the predictions at a certain time step. At each time step, a neuronal population samples the net excitatory input from all state vector – weight vector pairs $V(i,t) = x(i,t)\, w(i,t)$ (Montague, et al., 1996).

That is, it samples the net excitatory input from all cortical representations at timestep, considering the weight the input should have on the net excitation. It compares the excitatory input at the moment V(i,t) with the excitatory input one time step ago V(i,t-1). Summed over all domains ( $\Sigma_i$ V(i,t) – V(i,t-1)) this yields the net excitation V^.

$$\widehat{V}(t) \equiv \sum_i V(i,t) - V(i,t-1)$$

(Montague, et al., 1996).

If a salient or rewarding event (r(t)) occurs, this is separately registered by the neuronal population and the activity associated with the rewarding event is added to the difference of the excitatory input now and one time step ago. The product of the activity of the rewarding event and the difference between net excitatory activities between the time steps is the prediction error signal ($\delta$). This signal is the output of the neuronal population.

$\delta(t) = r(t) + V(t) - V(t-1)$(Montague, et al., 1996).

The prediction error signal is now sent back the cortical representations and changes the weight vectors. In clear terms: The prediction error output constantly adjusts the influence each cortical input region has on the activity in the neuronal population, which represents the predictions (Montague, et al., 1996). In a simple but widely used version of the model, only the weight of the last time step will be changed (Schultz et al., 1997). The new entry in the weight vector ($w_{new}$) concerning this last time step is the value of the prediction error multiplied with the state vector and a learning rate ($\eta$) plus the previous weight vector ($w_{prev}$). If the prediction error is large, the weight will be adjusted to a higher degree than if there is no prediction error.

$w(i, t-1)_{new} = w(i, t-1)_{prev} + \eta x(i, t-1)\delta(t)$ (Montague, et al., 1996).

If a salient or rewarding event occurs, the prediction error will be larger than if there was no such event and therefore change the influence the cortical region has on the activity of the neuronal population substantially. In the next iteration, the change of activity associated with the reward has propagated one time step backwards. This backwards propagation continues until the sensory state the full prediction error, but the reward is fully predicted and elicits no prediction errors.

## 4.3 Glossary

*Activity/activation, model induced*: Activity pattern corresponding to the expected activation encompassed in the generative model, results from back-projections.

*Activity/activation, unpredicted input*: Corresponds the activation elicited by the cause that is not predicted by the current generative model, is projected along forward-projections.

*Bias*: State signified by the presence of at least one strong, solid, or prepotent internal and one weaker internal model

*Caudate*: Basal ganglia nucleus, part of the striatum

*Cause*: External influence on perception that is represented in a generative model.

*Conflict*: A state of bias wherein external evidence, or response demands run against the stronger internal model.

*Direct pathway*: Transmission type in the basal ganglia, leads to a disinhibition of the thalamus and putatively to a fostering of the representation encoded currently in the direct pathway, including long term potentiation.

*Efference copy*: (originally) Signal that predicts the afferent input, which will result from the execution of motor command

*Entropy*: Measure of average amount of information/surprise derived from Information Theory

$$H(x_i) = \sum_{i-k} -p(x_i)\, x\, \ln p(x_i)$$

*Generative model*: In Bayesian terms: probability of the data given the hypothesis. In terms of predictive coding internal model of the probability of a neural activity pattern given the modelled cause.

*Globus pallidus*: Basal ganglia nucleus

*Habenula*: Epithalamus structure influencing dopaminergic midbrain

*Heteromodal association area*: (1) receive inputs from a number unimodal areas from different modalities, (2) neurons respond to input of different modalities, (3) lesions lead to multimodal behavioural deficits (Mesulam, 1998).

*Indirect pathway*: Transmission type in the basal ganglia, leads to an increased inhibition of the thalamus and putatively to a weakening of the representation encoded currently in the indirect pathway, including a lack of long term potentiation, and possible long term depression.

*Kullback-Leibler Divergence*: difference measure between two probability distributions

$$D(P;Q) = \sum P(data)\, log \frac{P(data)}{Q(data)}$$

*Likelihood*: The *posterior probability* before an observation has been made (therefore, it's not *posterior* to the current observation, but to past ones)

*Long Term Potentiation:* Plastic increase in synaptic strength, (here: D1 modulated)

*Long Term Depression:* Plastic decrease in synaptic strength, (here: D2 modulated)

*Prediction Error:* (in TD) Excitatory input now minus excitatory input one time step ago. Excitatory input can (and in most applications does) concerns mainly expectation of reward. $\delta(t) = r(t) + V(t) - V(t-1)$.

*Primary Association Area:* Cortical "input" area to different modalities, eg. striate cortex, Heschl's gyrus

*Putamen*: Basal ganglia nucleus, part of the striatum

*Shannon entropy*: cf *entropy*

*Striatum*: Caudate, Putamen, Nucleus accumbens (basal ganglia nuclei)

*Substantia nigra*: Dopaminergic midbrain area

*Subthalamic nucleus*: Basal ganglia nucleus

*Surprise*: Measure of information, ie., uncertainty reduction, of a stimulus $I(x_i) = -\ln p(x_i)$

*Temporal difference-algorithm*: Computational model, capable of learning temporal predictions; the algorithm learns when the achieved state diverges from the predicted state, ie. from prediction errors. Learning consists in adjustment of the weight the past state(s) has/ve on predictions.

*Unimodal association area:* (1) major source of input is primary association area or other unimodel areas of the same modality, (2) neurons respond preferentially or exclusively within one modality (3) lesions lead to deficits related to the specific modality (Mesulam, 1998)

*Ventral tegmental area*: Dopaminergic midbrain area

## 4.4 Abbreviations

ACC: anterior cingulate cortex

BA: Brodmann Area

D1 receptor: Dopamine receptor of the D1-family type (D1 & D5)

D2 receptor: Dopamine receptor of the D2-family type (D2, D3, & D4)

EEG: Electroencephalogram

ERN: Error related negativity

ERP: Event related potential

f-ERN: feedback related error related negativity

(f)MRI: (functional) magnetic resonance imaging

GPi: Globus pallidus interna (basal ganglia nucleus)

GPe: Globus pallidus externa (basal ganglia nucleus)

KBL: Kullback-Leibler Divergence

LTD: Long term depression

LTD: Long term potentiation

PM(v): (ventral) premotor cortex

SMA: Supplementary motor area

TD: temporal difference

VTA: Ventral tegmental area

SNc: Substantia nigra, pars compacta

SNr: Substantia nigra pars reticulata

STN: Subthalamic nucleus

V1: Primary (striate) visual cortex, Brodmann Area 17

V2: Secondary visual cortex, corresponds roughly to Brodmann Area 18 and 19

# 5 References

Albert, N. B., Peiris, Y., Cohen, G., Miall, R. C., & Praamstra, P. (2010). Interference effects from observed movement in Parkinson's disease. J Motor Behav, 42(2), 145-149.

Albin, R. L., Young, A. B., Penney, J. B., Roger, L. A., & Young, B. B. (1989). The functional anatomy of basal ganglia disorders. Movement Disord, 12(10).

Alexander, G. E., & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci, 13(7), 266-271.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci, 9, 357-381.

Alegre, M.C., Rodriguez-Oroz, M.C., Valencia, M., Perez-Alcazar, M., Guridi, J., Iriarte, J., Obeso, J. A., Artieda, J. (2010). Clinical Neurophysiol 414-425

Aosaki, T., Tsubokawa, H., Ishida, A., Watanabe, K., Graybiel, A. M., & Kimura, M. (1994). Responses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensorimotor conditioning . J Neurosci, 14 (6 ), 3969-3984.

Aziz-Zadeh, L., Koski, L., Zaidel, E., Mazziotta, J., & Iacoboni, M. (2006). Lateralization of the Human Mirror Neuron System. J Neurosci, 26(11), 2964-2970.

Baldi, P., & Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. Neural Networks, 23(5), 649-66.

Barbeau, A. (1970). Dopamine and disease . Can Med Assoc J, 103 (8), 824-832.

Beauchamp, M.S., Martin, A. (2007).Grounding Object Concepts in Action: Evidence from FMRI Studies of Tools. Cortex, 43(3) 461 - 468

Berke, J. D., & Hyman, S. E. (2000). Addiction, dopamine, and the molecular mechanisms of memory. Neuron, 25(3), 515-532.

Berkeley, G. (1709). An Essay towards a New Theory of Vision. Dublin.

Bernheimer, H., Birkmayer, W., Hornykiewicz, O., Jellinger, K., & Seitelberger, F. (1973). Brain dopamine and the syndromes of Parkinson and Huntington Clinical, morphological and neurochemical correlations. J Neurol Sci, 20(4), 415-455.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Res Rev, 28(3), 309-369.

Bischoff-Grethe, A., Crowley, M. G., & Arbib, M. A. (2002). Movement inhibition and next sensory state prediction in the basal ganglia. In A.M. Graybiel, M. R. DeLong, & S. T. Kitai (Eds.), The Basal Ganglia VI (pp. 267-278). New York: Kluwer Academic/Plenum Publishers.

Bjoerklund, A., & Dunnett, S. B. (2007). Dopamine neuron systems in the brain: an update. Trends Cogn Sci, 30(5), 2 194-201

Bolam, J. P., Brown, M. T. C., Moss, J., & Magill, P. J. (2009). Basal Ganglia : Internal Organization. Encyclopedia of Neuroscience, 97-104.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. Psychol Rev, 108(3), 624-652.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn Sci, 8(12), 539-546.

Bubic, A., von Cramon, D. Y., Jacobsen, T., Schroger, E., & Schubotz, R. I. (2009). Violation of expectation: neural correlates reflect bases of prediction. J Cogn Neurosci, 21(1), 155-168.

Buccino, G., Vogt, S., Ritzl, A., Fink, G. R., Zilles, K., Freund, H.-J., & Rizzolatti, G. (2004). Neural Circuits Underlying Imitation Learning of Hand Actions: An Event-Related fMRI Study. Neuron, 42(2), 323-334.

Bédard, P., Larochelle, L., Parent, A., & Poirier, L. J. (1969). The nigrostriatal pathway: A correlative study based on neuroanatomical and neurochemical criteria in the cat and the monkey. Exp Neurol, 25(3), 365-377.

Chouinard, G., & Jones, B. D. (1978). Schizophrenia as Dopamine-Deficiency Disease. The Lancet, 312 (8080) 99-100

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49 (12), 997-1003.

Crawford, T., Goodrich, S., Henderson, L., & Kennard, C. (1989). Predictive responses in Parkinson's disease: manual keypresses and saccadic eye movements to regular stimulus events. J Neurol Neurosurp Ps, 52(9), 1033-1042.

Cross, E. S., Kraemer, D. J. M., Hamilton, A. F. C., Kelley, W. M., & Grafton, S. T. (2009). Sensitivity of the Action Observation Network to Physical and Observational Learning. Cereb Cortex, 19(2), 315-326.

Csibra, G. (2007). Action mirroring and action interpretation: An alternative account. In P. Haggard, Y. Rosetti, & M. Kawato (Eds.), Sensorimotor Foundations of Higher Cognition. Attention and Performance XXII (pp. 435-459). Oxford: Oxford University Press.

Dagher, A., & Robbins, T. W. (2009). Personality, Addiction, Dopamine: Insights from Parkinson's Disease. Neuron, 61(4),502-510.

De Bruijn, E.R.A., de Lange, F.P., von Cramon, D.Y., & Ullsperger, M. (2009). When Errors are Rewarding. J Neurosci. 29(39), 12183-12186.

Decety, J., Grèzes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F., et al. (1997). Brain activity during observation of actions. Influence of action content and subject's strategy. Brain, 120 (10), 1763-1777.

den Ouden, H. E. M., Daunizeau, J., Roiser, J., Friston, K. J., Stephan, K. E., & Danizeau, J. (2010). Striatal prediction error modulates cortical coupling. J Neurosci, 30(9), 3210-9.

den Ouden, H. E. M., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2009). A dual role for prediction error in associative learning. Cereb cortex, 19(5), 1175-85.

Di Martino, A., Scheres, A., Margulies, D. S., Kelly, A. M. C., Uddin, L. Q., Shehzad, Z., Biswal, B., et al. (2008). Functional Connectivity of Human Striatum: A Resting State fMRI Study. Cereb Cortex, 18 (12 ), 2735-2747.

Donkers, F. C. L., Nieuwenhuis, S., & van Boxtel, G. J. M. (2005). Mediofrontal negativities in the absence of responding. Cogn Brain Res, 25(3), 777-87.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Networks, 12(7-8), 961-974.

Doya, K., & Ishii, S. (2007). A Probability Primer. In K. Doya, A. Pouget, & R. P. N. Rao (Eds.), The Bayesian brain: Probabilistic approaches to neural coding. Cambrige: MIT Press.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Dopamine-Mediated Stabilization of Delay-Period Activity in a Network Model of Prefrontal Cortex. J Neurophysiol, 83(3), 1733-1750.

Finlay, B. L., Schiller, P. H., & Volman, S. F. (1976). Quantitative studies of single-cell properties in monkey striate cortex. IV. Corticotectal cells. J Neurophysiol, 39(6), 1352-61.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci, 14(3), 119-130.

Flanagan, J Randall, & Johansson, R. S. (2003). Action plans used in action observation. Nature, 424(6950), 769-771.

Flowers, K. (1978). LACK OF PREDICTION IN THE MOTOR BEHAVIOUR OF PARKINSONISM. Brain, 101(1), 35-52.

Floyer-Lea, A., & Matthews, P. M. (2004). Changing Brain Networks for Visuomotor Control With Increased Movement Automaticity . J Neurophysiol, 92 (4 ), 2405-2412.

Frank, M. J. (2006). Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. Neural Networks, 19(8), 1120-36.

Frank, M.J., Claus, E. (2006). Anatomy of decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psych Rev, 113(2), 300-326.

Frank, M.J., Samanta, J., Moustafa, A.A., Sherman, S.J. (2007).Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. Science, 318 (5854), 1309-1312.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? Nat Rev Neurosci, 11(2), 127-138.

Friston, K. J. (2002). Functional integration and inference in the brain. Prog. Neurobiol., 68(2), 113-143.

Friston, K. J. (2005). A theory of cortical responses. Philos Trans R Soc Lond., 360, 815-136.

Friston, K. J. (2011). What Is Optimal about Motor Control? Neuron, 72(3), 488-498.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement Learning or Active Inference? PLoS ONE, 4(7), e6421. Retrieved from http://dx.doi.org/10.1371/journal.pone.0006421

Friston, K. J., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. Biol Cybern, 104(1-2), 137-160.

Garavan, H., Ross, T.J., Murphy, K., Roche, R.A.P., Stein, E.A. (2002). Dissociable Executive Functions in the Dynamic Control of Behavior, Inhibition, Error Detection, and Correction. NeuroImage, 17(4), 1820-1829.

Gardner, E. L., & Lowinson, J. H. (1993). Drug craving and positive/negative hedonic brain substrates activated by addicting drugs. Semin Neurosci, 5(5), 359-368.

Garner, W. R. (1975). Uncertainty and structure as psychological concepts. (p. ix, 369).Oxford:England, Wiley

Gaspar, P., Stepniewska, I., & Kaas, J. H. (1992). Topography and collateralization of the dopaminergic projections to motor and lateral prefrontal cortex in owl monkeys. J Comp Neurol, 325(1), 1-21.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A Neural System for Error Detection and Compensation . Psychol Sci, 4 (6 ), 385-390.

Gerfen, C. R., & Surmeier, D. J. (2011). Modulation of striatal projection systems by dopamine. Annu Rev Neurosci, 34, 441-66.

Ghahramani, Z., Wolpert, D. M., & Michale, I. J. (1997). Computational models of sensorimotor integration. In M. Pietro & S. Vittorio (Eds.), Advances in Psychology (Vol. 119, pp. 117-147). North-Holland.

Gibson, J. J. (1986). The ecological approach to visual perception. (p. 332). Boston: Houghton Mifflin.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychol Rev, 102(4), 684-704

Goldberg, G. (1985). Supplementary motor area structure and function: Review and hypotheses. Behav Brain Sci, 8(04), 567-588.

Graybiel, A M. (1998). The basal ganglia and chunking of action repertoires. Neurobiol Learn Mem, 70(1-2), 119-136.

Gruesser, O.-J. (1986). Interaction of Efferent and Afferent Signals in Visual Perception A History of Ideas and Experimental Paradigms. Acta Psychologica, 63, 3-21.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. Behav. Brain Sci, 27(3), 377-396.

Grèzes, J., & Decety, J. (2001). Functional anatomy of execution, mental simulation, observation, and verb generation of actions: A meta-analysis. Human Brain Mapping, 12(1), 1-19.

Haber, S. (2003). The primate basal ganglia: parallel and integrative networks. Journal of Chemical Neuroanatomy, 26(4), 317-330.

Hall, B. K. (1999). The paradoxical platypus, Biol Sci 49(3), 211-218.

Harrison, L. M., Duggins, a, & Friston, K. J. (2006). Encoding uncertainty in the hippocampus. Neural Networks, 19(5), 535-46.

Haruno, M., Wolpert, D. M., & Kawato, M. (2001). Mosaic model for sensorimotor learning and control. Neural Comput, 13(10), 2201-20.

Helmholtz, H. (1866). Physiologische Optik. In G. Karsten (Ed.), Allgemeine Encyklopaedie der Physik (pp. 28 - 32). Leipzig: Leopold Voss.

Herwig, A., & Waszak, F. (2009). Intention and attention in ideomotor learning. Q J Exp Psychol, 62(2), 219-27.

Herwig, A., Prinz, W., & Waszak, F. (2007). Two modes of sensorimotor integration in intention-based and stimulus-based actions. Q J Exp Psychol, 60(11), 1540-54.

Hikosaka, O., Nakamura, K., Sakai, K., & Nakahara, H. (2002). Central mechanisms of motor skill learning. Curr Opini Neurobiol, 12(2), 217-222.

Hikosaka, O., Sesack, S.R., Lecourtier, L., Shepard, P.D.(2008). Habenula: Crossroad between the basal ganglia and the limbic system. J Neurosci, 28(46), 11825-11829.

Holroyd, C.B. & Coles, M. (2002).The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. Psych Rev, 109(4), 679-709

Holroyd, C.B., Krigolson, O., Baker, R., Lee, S., Gibson, J.(2009). When is an error not a prediction error? An electrophysiological investigation. Cogn Affect Behav Neurosci, 9(1), 59-70

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G. H., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. Nat Neurosci, 7(5), 497-8. doi:10.1038/nn1238

Holroyd, C. B., Yeung, N., Coles, M. G. H., & Cohen, J. D. (2005). A mechanism for error detection in speeded response time tasks. J Exp Psychol, 134(2), 163-91.

Hong, S., Hikosaka, O.(2008). The Globus Pallidus Sends Reward-Related Signals to the Lateral Habenula. Neuron, 60(4), 720-729.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. Neurosci, 96(4), 651-656.

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. Wiley Interdisciplinary Reviews: Cognitive Science

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. J Neurophysiol, 28(2), 229-289.

Hughes, G., Yeung, N. (2011). Dissociable Correlates of Response Conflict and Error Awareness in Error-Related Brain Activity. Neuropsychologica, 49(3), 405-415.

Hurley, S. (2001). Perception And Action: Alternative Views. Synthese, 129(1), 3-40. Springer Netherlands.

Hurley, S. (2006). Active perception and perceiving action: The Shared Circuits Hypothesis. In T. Gendler & J. Hawthorne (Eds.), Perceptual Experience. Oxford University Press.

Iseki, K., Hanakawa, T., Shinozaki, J., Nankaku, M., & Fukuyama, H. (2008). Neural mechanisms involved in mental imagery and observation of gait. NeuroImage, 41(3), 1021-1031.

Itti, L., & Baldi, P. (2005). A Principled Approach to Detecting Surprising Events in Video. Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition, 1-7.

Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: a critique. Trends Cogn Sci, 9(1), 21-5.

James, W. (1890). CHAPTER XXVI. Will. The principles of psychology. Dover Punlications.

Jeannerod, M. (1995). Mental imagery in the motor context. Neuropsychologia, 33(11), 1419-1432.

Jocham, G., & Ullsperger, M. (2009). Neuropharmacology of performance monitoring. Neurosci Biobehav Rev, 33(1), 48-60.

Jueptner, M., & Weiller, C. (1998). A review of differences between basal ganglia and cerebellar control of movements as revealed by functional imaging studies. Brain, 121(8), 1437-1449.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), Punishment and aversive behavior. New York: Appleton-Century-Crofts.

Kelley, A. E. (2004).Memory and Addiction: Shared Neural Circuitry and Molecular Mechanisms. Neuron, 44(1), 161-179.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. Annu Rev Psych, 55, 271-304.

Keysers, C., & Perrett, D. I. (2004). Demystifying social cognition: a Hebbian perspective. Trends Cogn Sci, 8(11), 501-507.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. PLoS Comput Biol, 4(11), e1000209. doi:10.1371/journal.pcbi.1000209 [doi]

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). The mirror-neuron system: a Bayesian perspective. Neuroreport, 18(6), 619-623.

Kilner, J. M., Vargas, C., Duval, S., Blakemore, S.-J., & Sirigu, A. (2004). Motor activation prior to observation of a predicted movement. Nat Neurosci, 7(12), 1299-1301.

Kish, S. J., Shannak, K., & Hornykiewicz, O. (1988). Uneven Pattern of Dopamine Loss in the Striatum of Patients with Idiopathic Parkinson's Disease. New Engl J Med, 318(14), 876-880

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci, 27(12), 712-9.

Kornheiser, A. S. (1976). Adaptation to laterally displaced vision: A review. Psychol Bulletin, 83(5), 783-816.

Krieghoff, V., Waszak, F., Prinz, W., & Brass, M. (2011). Neural and behavioral correlates of intentional actions. Neuropsychologia, 49(5), 767-776.

Kühn, S., Seurinck, R., Fias, W., & Waszak, F. (2010). The Internal Anticipation of Sensory Action Effects: When Action Induces FFA and PPA Activity. Frontiers in human neuroscience, 4(), 54. doi:10.3389/fnhum.2010.00054

Laming, D.(2001). Statistical Information, Uncertainty, and Bayes' Theorem: Some Applications in Experimental Psychology. In: Benferhat, S., & Besnard, P. (Eds.):ECSQARU 2001.LNAI 2134, pp. 634-646, Heidelberg: Springer.

Lecourtier, L. & Kelly, P.H. (2007).A conductor hidden in the orchestra? Role of the habenular complex in monoamine transmission and cognition. Neurosci Biobehav Rev, 31(5), 658-672.

Le Moal, M., & Simon, H. (1991). Mesocorticolimbic dopaminergic network: functional and regulatory roles. Physiol Rev, 71(1), 155-234.

Levy, F., & Swanson, J. M. (2001). Timing, space and ADHD: the dopamine theory revisited. Australian and New Zealand Journal of Psychiatry, 35(4), 504-511.

Lindvall, O., Bjoerklund, A., & Divac, I. (1978). Organization of catecholamine neurons projecting to the frontal cortex in the rat. Brain Res, 142(1), 1-24.

Lindvall, O., Bjoerklund, A., & Skagerberg, G. (1984). Selective histochemical demonstration of dopamine terminal systems in rat di- and telecephalon: New evidence for dopaminergic innervation of hypothalamic neurosecretory nuclei. Brain Res, 306(1-2), 19-30.

Lindvall, O., Brundin, P., Widner, H., Rehncrona, S., Gustavii, B., Frackowiak, R., Leenders, K. L., et al. (1990). Grafts of fetal dopamine neurons survive and improve motor function in Parkinson's disease . Science , 247 (4942 ), 574-577.

Liu, K. P., Chan, C. C., Lee, T. M., & Hui-Chan, C. W. (2004). Mental imagery for promoting relearning for people after stroke: A randomized controlled trial. Archives of Physical Medicine and Rehabilitation, 85(9), 1403-1408.

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. J Neurophysiol, 67(1), 145-163.

Luce, R.D. (2003).Whatever happened to information theory in psychology? Rev General Psychology, 7(2), 183-188.

Matsumoto, M., Hikosaka, O., (2007). Lateral Habenula as a Source of Negative Reward Signals in Dopamine Neurons. Nature, 447(7148), 1111-1115.

Mesulam, M.(1998). From Sensation to Cognition. Brain, 121(6). 1013-1052.

Mehta, M., Neuronal Dynamics of Predictive Coding, Neuroscientist, 7(6), 490-495.

McHaffie, J. G., Stanford, T. R., Stein, B. E., Coizet, V., & Redgrave, P. (2005). Subcortical loops through the basal ganglia. Trends Neurosci, 28(8), 401-407.

Meyer, A., & Hierons, R. (1964). A note on Thomas Willis' views on the corpus striatum and the internal capsule. J Neurol Sci, 1(6), 547-554.

Miall, R. C. (2003). Connecting mirror neurons and forward models. Neuroreport, 14(17), 2135-7.

Mink, J. M. (1996). The Basal Ganglia: Focused Selection and IInhibition of Competing Motor Programs. Progr Neurobiol, 50(4), 381-425.

Missale, C., Nash, S. R., Robinson, S. W., Jaber, M., & Caron, M. G. (1998). Dopamine receptors: from structure to function. Physiol Rev, 78(1), 189-225

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci, 16(5), 1936-1947.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. Nat Neurosci, 9(8), 1057-1063.

Mueller, J. (1838). Introduction to Handbuch der Physiologie des Menschen. Handbuch der Physiologie des Menschen. Koblenz.

Munzert, J., Zentgraf, K., Stark, R., & Vaitl, D. (2008). Neural activation in cognitive motor processes: comparing motor imagery and observation of gymnastic movements. Exp Brain Res, 188(3), 437-444.

Nyberg, L., Eriksson, J., Larsson, A., & Marklund, P. (2006). Learning by doing versus learning by thinking: An fMRI study of motor and mental training. Neuropsychologia, 44(5), 711-717.

O'Doherty, J. P., Buchanan, T. W., Seymour, B., & Dolan, R. J. (2006). Predictive Neural Coding of Reward Preference Involves Dissociable Responses in Human Ventral Midbrain and Ventral Striatum. Neuron, 49(1), 157-166

O'Doherty, J. P., Dayan, P., Friston, K. J., Critchley, H., & Dolan, R. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. Neuron, 38(2), 329-337.

O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K. J., & Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. Science, 304(5669), 452-454.

Olivera, F.T., McDonald, J.J., Goodman, D.(2007). Performance monitoring in the anterior cingulate cortex is not all error-related: expectancy deviation and the representation of action-oytcome associations. J Cogn Neurosci, 19(12), 1994-2004.

Pan, W. X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J Neurosci, 25(26), 6235-6242

Parent, A, & Hazrati, L.-N. (1995a). Functional anatomy of the basal ganglia. I. Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. Brain Res Rev, 20, 91-127.

Parent, A, & Hazrati, L.-N. (1995b). Functional anatomy of the basal ganglia. II. The place of subthalamic nucleus and external pallidum in basal ganglia circuitry. Brain Res Rev, 20, 128-154.

Picard, N., & Strick, P. L. (2001). Imaging the premotor areas. Curr Opin Neurobiol, 11(6), 663-672.

Potts, G.F., Martin, L.E., Kamp, S., Donchin, E. (2011). Neural response to action and reward prediction errors: Comparing the error-related negativity to behavioral errors and the feedback-related negativity to reward prediction violations. Psychophysiol, 48(2), 218-228.

Ramsey, R., & Hamilton, A. F. C. (2010). How does your own knowledge influence the perception of another person's action in the human brain? Soc Cogn Affect Neurosci.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci, 2(1), 79-87.

Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? Nat Rev Neurosci, 7(12), 967-975

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). THE BASAL GANGLIA : A VERTEBRATE SOLUTION TO THE SELECTION PROBLEM ? Science, 89(4), 1009-1023.

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

Reynolds, J. N., & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. Neural Networks, 15(4-6), 507-21.

Rose, J., Schiffer, A.-M., Dittrich, L., & Gunturkun, O. (2010). The role of dopamine in maintenance and distractability of attention in the "prefrontal cortex" of pigeons. J. Neurosci.

Rushworth, M.F.S., Behrens, T.E.J.(2008). Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci, 11(4), 389-397.

Saint-Cyr, J. A. (2003). Frontal-striatal circuit functions: context, sequence, and consequence. J Int Neuropsychol Soc, 9(1), 103-127.

Schiffer, A.-M., & Schubotz, R. I. (2011). Caudate nucleus signals for breaches of expectation in a movement observation paradigm. Frontiers in human neuroscience, 5, doi:10.3389/fnhum.2011.00038

Schiffer, A.-M., Ahlheim, C., Ulrichs, K., & Schubotz, R.I. (in Press). Neural Changes When Actions Change: Adaptation of Strong and Weak Adaptations. Human Brain Mapping.

Schiffer, A.-M., Ahlheim, C., Wurm, M. F., & Schubotz, R. I. (submitted). Surprised at all the Entropy: Hippocampal, Caudate and Midbrain Contributions to Learning from Prediction Errors.

Schiller, P. H., Finlay, B. L., & Volman, S. F. (1976). Quantitative studies of single-cell properties in monkey striate cortex. V. Multivariate statistical analyses and models. J Neurophysiol, 39(6), 1288-319.

Schmidt, R. (2005). Exploration and extension of temporal-difference models of midbrain dopamine cell firing. University of Otago.

Schubotz, R. I. (2007). Prediction of external events with our motor system: towards a new framework. Trends Cogn Sci, 11(5), 211-218.

Schubotz, R. I., & von Cramon, D. Y. (2003). Functional-anatomical concepts of human premotor cortex: evidence from fMRI and PET studies. NeuroImage, 20, S120-S131

Schubotz, R. I., & von Cramon, D. Y. (2004). Sequences of Abstract Nonbiological Stimuli Share Ventral Premotor Cortex with Action Observation and Imagery. J Neurosci, 24(24), 5467-5474.

Schubotz, R. I., Korb, F. M., Schiffer, A.-M., Stadler, W., & von Cramon, D. Y. (submitted). The fraction of an action is more than a movement: Neural signatures of event segmentation in fMRI.

Schultz, W. (2000). Multiple reward signals in the brain. Nat Rev Neurosci (3)

Schultz, W. (2007). Multiple dopamine functions at different time courses. Annu Rev Neurosci, 30, 259-288.

Schultz, W., & Dickinson, A. (2000). Neuronal Coding of Prediction Errors. Annu Rev Neurosci, 23(1), 473-500.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task . J Neurosci, 13 (3 ), 900-913

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. Science (New York, N.Y.), 275(5306), 1593-1599.

Selemon, L. D., & Goldman-Rakic, P. S. (1985). Longitudinal topography and interdigitation of corticostriatal projections in the rhesus monkey. J. Neurosci., 5(3), 776-794.

Shane, M.S., Stevens, M., Harenski, C.L., & Kiehl, K.A.(2008). Neural correlates of the processing of another's mistakes: a possible underpinning for social and observational learning. NeuroImage, 42(1), 450-459.

Shannon, C. E., & Weaver, W. (1949). The Mathematical Theory of Information. Urbana, Illinois: University of Illinois Press.

Smith, Y., Bevan, M. D., Shink, E., & Bolam, J. P. (1998). MICROCIRCUITRY OF THE DIRECT AND INDIRECT PATHWAYS OF THE BASAL GANGLIA. Science, 86(2), 353-387.

Sperry, R. W. (1950). NEURAL BASIS OF THE SPONTANEOUS OPTOKINETIC RESPONSE PRODUCED BY VISUAL INVERSION. J Comp Psychol, 43(6),482-489.

Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., & Friston, K. J. (2005). Information theory, novelty and hippocampal responses: unpredicted or unpredictable? Neural Networks, 18(3), 225-30.

Sturrock, A., & Leavitt, B. R. (2010). The Clinical and Genetic Features of Huntington Disease. J Geriatr Psychiatry Neurol, 5, Published Online before print.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006a). Predictive codes for forthcoming perception in the frontal cortex. Science , 314(5803), 1311-4.

Summerfield, C., Egner, T., Mangels, J., Hirsch, J., (2006b). Mistaking a house for a face: neural correlates of misperception in healthy humans. Cereb Cortex. 14(4), 500-508.

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci, 11(9), 1004-1006.

Suri, R. E. (2002). TD models of reward predictive responses in dopamine neurons. Neural Networks, 15(4-6), 523-533.

Sutton, R. S., & Barto, A. G. (1990). Chapter 12. Time-Derivative Models of Pavlovian Reinforcement, Learning and Computational Neuroscience: Foundations of Adaptive Networks, Gabriel, M. & Moore, J. (Eds), pp. 497-537, MIT Press

Takakusaki, K., Saitoh, K., Harada, H., & Kashiwayanagi, M. (2004). Role of basal ganglia–brainstem pathways in the control of motor behaviors. Neurosci Res, 50(2), 137-151.

Thorndike, E.(1927). The Law of Effect. Am J Psychol, 39(1), 212-222.

Turk-Browne, N. B., Scholl, B.K, Johnson, M.K., & Chun, M.M.(2010). Implicit Perceptual Anticipation Triggered by Statistical Learning. J Neurosci, 30(33), 11177-11187.

Ullsperger, M., & von Cramon, D. Y. (2003). Error monitoring using external feedback: specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging. J Neurosci, 23(10), 4308-4314.

Ullsperger, M., & von Cramon, D. Y. (2004). Neuroimaging of Performance Monitoring: Error Detection and Beyond. Brain, 40(6), 593-604.

Ullsperger, M., & von Cramon, D. Y. (2006). The role of intact frontostriatal circuits in error processing. J Cogn Neurosci, 18(4), 651-664.

Van de Veen, A., Schouten, B.(2010). A minumum relative entropy principle for AGI. Proceedings of the Third Conference on Artificial General Intelligence.

Van Schie, H.T., Mars, R.B., Coles, M.G.H., & Bekkering, H.(2004). Modulation of activity in medial frontal and motor cortices during error observation. Nat Neurosci, 7(5), 549-554.

Von Holst, E., & Mittelstaedt, H. (1950). Das Reafferenzprinzip . ( Wechlselwirkungen zwischen Zentralnervensystem und Peripherie.). Die Naturwissenschaften, 464 - 475.

VonSattel, J.-P., Myers, R. H., Stevens, T. J., Ferrante, R. J., Bird, E. D., & Richardson, E. P. J. (1985). Neuropathological Classification of Huntington's Disease. J Neuropath & Exp Neurol, 44(6).

Waszak, F., Wascher, E., Keller, P., Koch, I., Aschersleben, G., Rosenbaum, D. A., & Prinz, W. (2005). Intention-based and stimulus-based mechanisms in action selection. Exp Brain Res,162(3), 346-56.

Watson, J. B. (1913). Psychology as the Behaviorist Views it. Psychol Rev, 20(158-177).

Wickens, J. R.(2008) Towards an Anatomy of Disappointment: Reward-Related Signals from the Globus Pallidus. Neuron, 60(4), 530-531.

Wickens, J. R., Horvitz, J. C., Costa, R. M., & Killcross, S. (2007). Dopaminergic mechanisms in actions and habits. J Neurosci 27(31), 8181-3.

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. Nat Neurosci, 3, 1212-1217.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. Neural Networks, 11(7-8), 1317-29.

Wolpert, D. M., & Miall, R. C. (1996). Forward Models for Physiological Motor Control. Neural Networks, 9(8), 1265-1279.

Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of sensorimotor learning. Nat Rev Neurosci, 12,739-751

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. Phil Trans Royal Soc Lond B, 358(1431), 593-602.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. Science, 269(5232), 1880-1882.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. Psychol Rev, 111(4), 931-959.

Yágüez, L., Canavan, A. G. M., Lange, H. W., & Hömberg, V. (1999). Motor learning by imagery is differentially affected in Parkinson's and Huntington's diseases. Behav Brain Res, 102(1-2), 115-127.

Yágüez, L., Nagel, D., Hoffman, H., Canavan, A. G. M., Wist, E., & Hömberg, V. (1998). A mental route to motor learning: Improving trajectorial kinematics through imagery training. Behav Brain Res, 90(1), 95-106.

Zacks, J. M., & Swallow, K. M. (2007). Event Segmentation. Curr Dir Psychol Sci, 16(2), 80-84.

Zacks, J. M., Braver, T. S., Sheridan, M. a, Donaldson, D. I., Snyder, a Z., Ollinger, J. M., Buckner, R. L., et al. (2001). Human brain activity time-locked to perceptual event boundaries. Nat Neurosci, 4(6), 651-5.

Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction Error Associated with the Perceptual Segmentation of Naturalistic Events. J Cogn Neurosci, 23(12), 4057-4066. MIT Press.

## 6 List of Figures:

Figure 1: The encompassed figures were added as Supplementary Material to the submission to allow an unbiased evaluation of the handdrawn ROIs. They are presented separately in this graph as they do not appear in the ‚Authors’ proof’ that is encompassed above. S-Figure 1: Hippocampus y = -29; S-Figure 2: Hippocampus x = -28; S-Figure 3: Caudate y = 9; S-Figure 4: Caudate x= 12; S-Figure 5: Habenula z = 3; S-Figure 6: Substantia nigra y = -20.

.

## Curriculum Vitae

Personal

| | |
|---|---|
| Name: | Helga Anne-Marike Schiffer-Maraun |
| Born: | July 12th 1985, Cologne |
| Marital status: | Married |

Education

| | |
|---|---|
| Since 10/2011 | Postgraduate student at the Westfaelische Wilhelms-Universitaet, Faculty of Psychology |
| 09/2010- 09/2011 | Maastricht University, Faculty of Psychology and Neuroscience, M.Sc. Neuropsycholgy |
| 09/2005-02/2009 | Ruhr-University Bochum, Faculty of Psychology, B.Sc. Psychology |

Research

| | |
|---|---|
| 05/2009 – 03/2012 | Max Planck Institute for Neurological Research, Cologne, Motor Cognition Group |
| 03/2008 – 02/2009 | Ruhr-University Bochum, Institute of Cognitive Neuroscience, Department of Biopsychology |
| 10/2007- 1/2008 | Otago University, Dunedin, Memory and Cognition Lab |

Employment

| | |
|---|---|
| 02/2009 – 07/2009 | Max Planck Institute for Neurological Research, Cologne, Motor Cognition Group, Student assistant |
| 05/2009 – 09/2009 | Ruhr-University Bochum, Institute of Cognitive Neuroscience, Department of Biopsychology, Scientific assistant |
| 07/2006 – 10/2006 | Ruhr-University Bochum, Institute of Cognitive Neuroscience, Department of Cognitive Psychology, Student assistant |
| 09/2005 – 1/2009 | dbb akademie, Bonn, Student assistant |

## Declaration

# Eidesstattliche Versicherungen

Hiermit versichere ich, Helga Anne-Marike Schiffer-Maraun, dass ich:

1. Nicht wegen eines Verbrechens zu dem ich meine wissenschaftliche Qualifikation missbraucht habe, verurteilt worden bin.

2. Keine frueheren Promotionsversuche unternommen habe.

3. Die Dissertation nicht bereits anderweitig als Pruefungsarbeit vorgelegt habe.

4. Die vorgelegte Dissertation selbst und ohne unerlaubte Hilfe angefertigt habe, sowie alle in Anspruch genommenen Quellen und Hilfsmittel in der Dissertation angegeben habe.

5. Die Beschreibung der experimentellen Arbeiten, wie in der vorgelegten Dissertation gekennzeichnet, auf drei folgenden wissenschaftlichen Abhandlungen basiert und ich in allen als verantwortlicher Autor (”corresponding author”) gekennzeichnet bin, da ich hauptverantwortlich an der Entwicklung der Fragestellung und des experimentellen Designs, der Datenerhebung und Datenauswertung, sowie der Interpretation und Verfassung der hier genannten Manuskripte beteiligt war:

Schiffer, A.-M., & Schubotz, R. I. (2011). Caudate nucleus signals for breaches of expectation in a movement observation paradigm. Frontiers in human neuroscience, 5, 38.

Schiffer, A.-M., & Schubotz, R.I. (in press). Neural Changes When Actions Change: Adapatation of Strong and Weak Expectations. Human Brain Mapping.

Schiffer, A.-M., Ahlheim, C., Wurm, M. F., & Schubotz, R. I. (submitted). Surprised at all the Entropy: Hippocampal, Caudate and Midbrain Contributions to Learning from Prediction Errors.

Koeln, den ___________________ Unterschrift_______________________________

Addendum:
Die Manuskripte der drei wissenschaftlichen Abhandlungen sind in der durch den Review-Prozess bedingten, jeweils mir neuesten vorliegenden Form abgedruckt. Im Fall der Abhandlung: “Neural Changes When Actions Change: Adapatation of Strong and Weak Expectations.” ist dies der “Authors’ proof” der bereits akzeptierten Studie; die Korrekturen sind eingereicht. Im Fall der Abhandlung: “Surprised at all the Entropy: Hippocampal, Caudate and Midbrain Contributions to Learning from Prediction Errors.” liegt der der “Authors’ proof” der Einreichung am 24.12.2011 vor. Da in diesem Manuskript das “Supplementary Material” nicht erscheint, sind diese Grafiken zusaetzlich, auf der auf das Manuskript folgenden Seite 102 abgedruckt.

Koeln, den ____________________ Unterschrift________________________________