From Actions to Agents:
Value Representation in Frontal Cortex

Marco K Wittmann

Department of Experimental Psychology

Wadham College

Trinity term 2016

A dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the University of Oxford

From Actions to Agents:
Value Representation in Frontal Cortex

Abstract

In this thesis, I investigated computational and neural mechanisms underlying foraging-related behavior in humans. A consideration of the ecological constraints under which mammalian behavior first evolved guided my investigation of learning and decision-making in frontal cortex. When engaged in foraging, animals have to figure out how profitable their actions are and whether it is better to continue foraging in their current environment or to switch to an alternative. They have to track the reward income of their actions over time and also take into account that the actions of other foraging animals have a direct influence on their reward income (Chapter 1). Based on these observations, I looked at similar types of behavior in humans using computational modeling and functional magnetic resonance imaging. I studied how people evaluate the profitability of their actions over time (Chapter 2). Dorsal anterior cingulate cortex carried a detailed representation of the value of the current foraging action, which was influenced by reward memories with different time constants (Chapter 3). Tracking the reward income of one's actions is not only important in order to learn about the profitability of the environment, it can also inform estimates of one's own and other people's abilities. Ability estimates can be used in a direct way to predict the reward outcome that the actions of one's own self and others will have. I found that people learn about their own and others' abilities in a rational manner but also that ability estimates of self and other were partly confused with each other, depending on whether subjects cooperated or competed with each other (Chapter 4). The confusion effect is reflected in Brodmann area 9 activity indicating that area 9 integrates self and other related information. Perigenual anterior cingulate tracked the ability estimates for oneself, suggesting that it might compute the success expectation of an action independent of particular features of the environment (Chapter 5). In sum, different subregions of medial frontal cortex carried different types of action-related value representations that can guide decision-making.

Marco K Wittmann
Wadham College, Trinity term 2016
Approximate word count: 50.000 words
Submitted for the degree of Doctor of Philosophy

This work is dedicated to my parents, who made all this possible.

Table of Contents

CHAPTER 1: General introduction **6**

1.1. Theoretical framework and research questions *6*

1.1.1. Behavioral ecology and the brain 6

1.1.2. Patchy environments and the marginal value theorem 8

1.1.3. Competitive foraging and ideal free distribution 9

1.1.4. Summary 12

1.2. Methods in cognitive neuroscience *14*

1.2.1. In vivo electrophysiological recordings 14

1.2.2. Lesion and inactivation studies 15

1.2.3. Neuroimaging 16

1.2.4. Computational modeling 18

1.3. Medial frontal cortex: Anatomy *20*

1.3.1. Dorsal anterior cingulate cortex 20

1.3.2. Perigenual anterior cingulate cortex 23

1.3.3. Area 9 24

1.3.4. Ventromedial prefrontal cortex 27

1.4. Medial frontal cortex: Function *30*

1.4.1. Dorsal anterior cingulate cortex 30

1.4.2. Area 9 38

1.4.3. Perigenual anterior cingulate cortex 44

1.4.4. Ventromedial prefrontal cortex 50

1.4.5. Summary 51

CHAPTER 2: Tracking the success of actions over time **54**

2.1. Introduction *54*

2.2. Methods *56*

2.2.1. Subjects 56

2.2.2. Experimental design 56

2.2.3. Training sessions 60

2.2.4. Main reinforcement learning models 62

2.2.5. Supplementary reinforcement learning models 67

2.2.6. Behavioral and reinforcement learning model analysis 68

2.3. Results *71*

2.3.1. Opposing effects of recent and past rewards on choice 71

2.3.2. Estimating prediction errors enables reward trend guided choice 75

2.3.3. Supplementary RL analyses 79

2.4. Discussion *90*

CHAPTER 3: Multiple time-linked reward signals in dACC **92**

3.1. Introduction *92*

3.2. Methods *94*

3.2.1. Imaging data acquisition and preprocessing 94

3.2.2. FMRI whole-brain analyses 94

3.2.3. FMRI region of interest analyses 96

3.2.4. Supplementary fMRI analyses 98

3.3. Results *101*

3.3.1. dACC contrasts recent and past rewards 101

3.3.2. Prediction errors and choice signals in medial prefrontal cortex 106

3.4. Discussion *112*

CHAPTER 4: Predicting actions of oneself and others **115**

4.1. Introduction *115*

4.2. Methods *117*

4.2.1. Subjects 117

4.2.2. Overview over experiment structure 117

4.2.3. Detailed experimental design and schedule 121

4.2.4. Ability ratings 125

4.2.5. Feedback 127

4.2.6. Experimental procedure, instructions and debriefing questionnaire 130

4.2.7. Minigames 135

4.2.8. Reinforcement learning model 138

4.2.9. Behavioral analyses 142

4.3. Results *147*

4.3.1. Appropriate ability estimation and self-other-mergence 147

4.3.2. Detailed interaction GLMs 150

4.3.3. Behavioral adjustments driven by self-other-mergence 153

4.4. Discussion *156*

CHAPTER 5: Neural correlates of ability estimation **158**

5.1. Introduction *158*

5.2. Methods *161*

5.2.1. FMRI data acquisition and preprocessing 161

5.2.2. FMRI whole-brain analyses 161

5.2.3. FMRI region of interest analyses 165

5.3. Results *168*

5.3.1. Appropriate ability estimation and self-other-mergence 168

5.3.2. Supplementary results related to the decision phase 173

5.3.3. Behavioral adjustments linked to self-other-mergence in area 9 177

5.4. Discussion *180*

CHAPTER 6: General discussion **184**

Bibliography **189**

CHAPTER 1: General introduction

1.1. Theoretical framework and research questions

1.1.1. *Behavioral ecology and the brain*

The theoretical framework that guided the work on this thesis stems from behavioral ecology, in particular optimal foraging theory (Stephens and Krebs, 1986). According to this perspective, adaptive behavior is a product of evolution. Behavior is shaped over generations by natural selection under the constraints of the environment. These constraints are manifold and therefore pose separate evolutionary challenges to animals that need to be addressed in order to survive and reproduce. Behavioral ecology seeks to identify these challenges as they exist in natural environments and to summarize their general structure. Within this framework, optimal foraging theory is notably a normative approach (Kacelnik and Bateson, 1997) that seeks to determine the types of behavior that are necessary to maximize the longer-term average reward rate or fitness of an animal, given evolutionary challenges and the structure of the environment. It rests on the assumption that optimal behavior is likely to have been selected over less optimal behavior during evolution. Importantly, the notion of optimality is used here in reference to a specific ecological problem, but does not imply animals or humans behave optimally in any arbitrary task. Predictions derived from the optimal solution are subsequently compared with the actual behavior of animals and this often uncovers that behavioral strategies across different species are surprisingly optimally tuned to their environments. Note, however, that although different animals may solve an ecological problem in a similar way, this does not necessarily mean that they use the same computational mechanisms for solving it. Critical variables that influence an animals' fitness comprise the rewards it harvests by foraging, the direct and indirect costs of its activities, such as effort and time, and the competition with other foraging animals.

By contrast, in cognitive neuroscience, we are often looking for evidence of neural computations that drive behavior. Imaging brain activity during task performance may reveal something about those mechanisms. However, a consideration of the ecological constraints under which mammalian behavior first evolved might help both in the design of cognitive tasks and in the interpretation of the neural activity that is identified.

Foraging frameworks have been employed successfully in cognitive neuroscience on several occasions recently (Genovesio et al., 2014; Hayden et al., 2011b; Kolling et al., 2012; Pearson et al., 2014; Sugrue et al., 2004). This approach complements others such as neuroeconomics (Glimcher, 2004) which has introduced important concepts such as framing, subjective value, and risk from economics and demonstrated their relevance and importance for cognitive neuroscience (De Martino et al., 2006; Hsu et al., 2005; Kable and Glimcher, 2007; Tom et al., 2007).

The research performed for this thesis was guided by behavioral ecology in two ways. First, I focused on how humans can exploit the temporal dynamics of natural environments to inform choice and, second, I investigated how they engage in interactions with other conspecifics. In detail, the lines of investigation were the following two:

1. A general feature of ecological foraging is that reward rates are auto-correlated over time; they often change in a slow and predictable fashion, for instance when a source of food slowly depletes while the animal harvests from it. How do humans exploit this feature when choosing to disengage from a currently exploited environment and engage in a new one? (chapters 2 and 3)

2. The presence of other humans or animals is a fundamental constituent of our environments and they have an impact on whether our own reward rates increase or decrease. What types of self and other representation does the brain possess and how do we compute relationships between self and others? (chapters 4 and 5)

The next two sections give more detailed background to the two lines of investigation.

1.1.2. *Patchy environments and the marginal value theorem*

One of the most influential theories in optimal foraging is Charnov's 'marginal value theorem' (MVT) (Charnov, 1976). It conceptualizes animals' environments and their behavior in a very particular way. It envisages the environment as organized in a series of patches, i.e. separable sub-environments where the animal spends time foraging (Fig.1). Foraging is a continuous activity in which the animal invests time and energy to harvest rewards that it needs to survive. The animal's goal is to maximize its long-term average reward rate, i.e. the rewards it harvests given the cost it pays. It cannot control the rate of return that a patch yields, i.e. the reward rate. Eventually, when foraging in a patch, the reward rate will decrease and asymptote to zero when the patch is depleted. However, before that, reward rates may not always be decreasing. They may also increase, for instance if the predator's arrival in a patch scares prey into revealing their presence or if

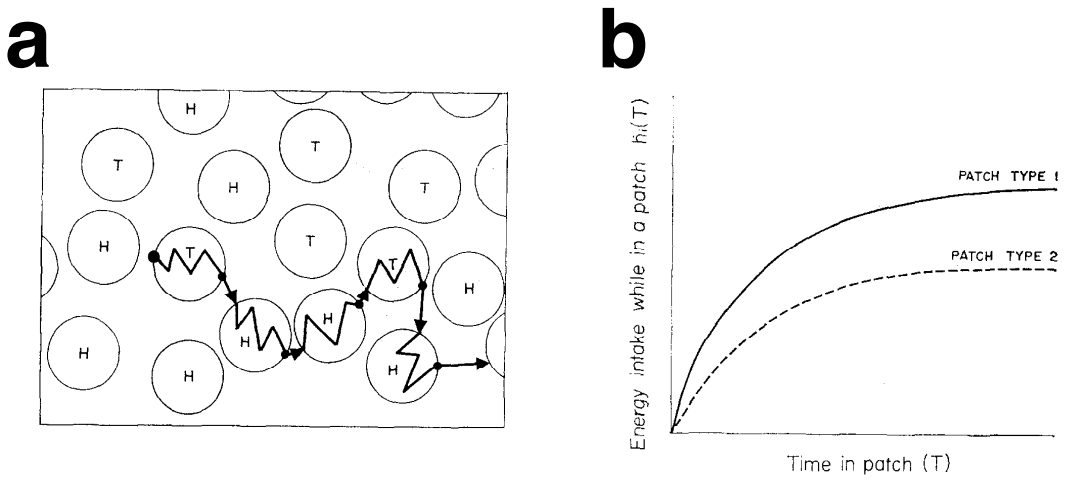


Figure 1. Patch foraging as continuous behavior. *(**a**) Environment with two patch types, H and T, shown as circles. The thick black line describes the foraging path of a predator spending time in patches to forage and travelling between patches. (**b**) Rate of return in two example foraging patches. Note that the graph shows the cumulative energy intake rather than the reward rates in the patches. The latter would be the first derivative of the curves shown. [Panels reprinted from (Charnov, 1976) with permission from Elsevier]*

initial investments are required to harvest fruits in a difficult-to-reach location, which then become more readily available (Charnov, 1976; Stephens and Krebs, 1986).

The critical variable that the animals has control over is whether it stays in the current patch and continues foraging or leaves the patch to search for an alternative. In this type of decision, there is an asymmetry between the two options. While the choice of staying is usually better defined and results in immediate reward, leaving requires the animal to delay reward receipt by travelling to another patch with an unclear reward distribution. However, this can still be beneficial in the long-term if the goal of the animal is to maximize the average reward income. MVT's prediction is that the animals will disengage from a patch if the reward rate drops below the average reward rate of the environment (i.e. the average of all patches combined). Note, however, that this prediction rests on the assumption that the animal has previously identified that the patch is about to deplete rather than is about to increase in value. If this were not the case then animals would never engage in patches that were initially poor, poorer than average, and only subsequently increased in value. The leave-stay decision is a decision about which action to take rather than, for example, which stimulus to pick out of an offer set as many paradigms studied in behavioral economics. It is also inherently connected to the notion of time and changes in reward rates as the animal has to continuously track the value of its actions to effectively engage and disengage in a patch.

1.1.3. *Competitive foraging and ideal free distribution*

The presence of other animals in a patch has a fundamental relevance for foraging behavior. Conspecifics foraging in the same patch represent competition and have a negative influence on the reward income of an animal. The concept of 'ideal free distribution' (Fretwell and Lucas, 1969; Wiens, 1976) suggests that animals distribute themselves over patches so that the reward rates of all individual animals are

approximately the same. Note that animals may differ in their competitive ability; they might be bigger or smaller, requiring more or less resources (Ruxton and Humphries, 1999). This can be factored in when calculating the ideal distribution ratios. Ideal free distribution assumes that patches differ in their quality, i.e. in how profitable they are when investing time to forage in them. At the same time conspecifics may forage in the same patch and this will reduce the rate of return for the individual animal. As in the MVT, animals can control whether they want to continue foraging in a patch or if they switch to an alternative. If the overall number of foraging animals is low, all of them may choose to forage in the most profitable patch. However, as competition increases the reward rate of an animal will decrease and eventually fall below the reward rate that a less profitable - but also less competitive - patch would offer. Then, animals begin to switch to the less profitable patch until an equilibrium emerges in which the reward income is matched for all animals. A famous test of this theory was conducted by Harper in which bread was thrown in a duck pond at different locations and the ratio of ducks approaching the different bread delivery sites matched the ratio of bread delivered (Harper, 1982) (Fig.2).

As previously mentioned, optimal foraging theory makes no predictions about the specific mechanisms that underlie the adaptive behavioral responses of the animals and it is likely that these mechanisms differ between species. The response to social competition, predicted by the ideal free distribution theory, may be implemented in different ways. For example, the foraging animal may encounter less prey in the most profitable environment as it has already been depleted by other conspecifics and therefore it may leave to a less profitable habitat with less competition. In this case, the effect of social competition on the animal's reward income may be indirect and mediated by the reduced number of prey items. However, a more elegant mechanism would be if the animal was aware of the rate of reward with which the other animals forage and the rate of reward with which it itself forages. Such a representation of an agents' reward rate would allow the animal to

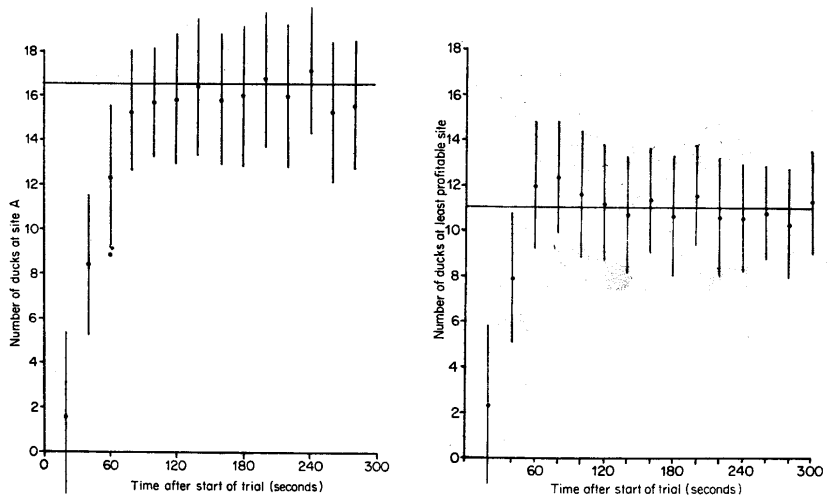


Figure 2. Ideal free distribution. *Data from the 'duck pond' experiment by Harper (Harper, 1982). Bread was thrown into a duck pond with different ratios. Duck number at different bread distribution sites plotted over time. The horizontal black lines indicate the optimal number of ducks per site according to the ideal free distribution. Left: Ratio of 1:1 distribution of bread per site. Right: Ratio of 2:1 distribution of bread per site, shown is the less profitable site. [Reprinted from (Harper, 1982) with permission from Elsevier]*

predict in advance how others affect its own reward income and would enable it to switch away from overpopulated patches even before its own reward rate falls below a critical level. In addition, such a representation of competitive ability might be informed by the relationship between the animals. In cooperative hunting, as for example practiced by a number of primate species including chimpanzees and humans, animals in the hunting party initially cooperate to hunt down prey, but later have to compete amongst themselves for the prey; the relationship in the hunting party flips from cooperation to competition and an ideal group size can be calculated that reflects the balance between the two forms of engagement: bigger group sizes lead to more efficient hunts and the capacity to hunt down stronger prey items; however, the food intake per prey item per hunting animal is reduced in bigger groups (Creel and Creel, 1995; Pulliam and Caraco, 1984).

In sum, other animals have a vital relevance for the evolutionary fitness of an animal. If an animal knew about its own and the other's rates of reward, as well as the

cooperative and competitive relationship to them, then this might greatly improve an animal's long-term reward income by allowing the animal to predict others' impact on the reward environment.

1.1.4. *Summary*

Behavioral ecology helps to identify the types of behavior that are fundamental for survival in natural environments. Foraging animals have to continuously track the value of their actions as their reward rates change systematically over time. Their choice is often whether to continue with the present action or to disengage and switch to an alternative option. In both cases, the focus of the choice is an activity rather than the choice between specific stimuli. The main question is how the animal allocates its time. Simple threshold-based choice strategies (e.g. *if* "reward rate" < X, *then* "leave patch") are certainly of central importance in this. However, such strategies might be enriched by temporally more detailed representations of the reward environment. The animal might not only be aware of the value of its current actions, but also the values of its previous actions and use this knowledge to compute the overall temporal structure of the reward environment. One advantage the animal could gain from this is an estimate of the direction of change of reward rates. This might allow the decision-maker for instance to prospectively disengage from depleting patches even before experiencing that the reward rate drops below a threshold, optimizing the time and effort allocated in reward patches in the long-term.

It is not only important to take into account one's own actions when foraging, but also the actions of other animals. They have a direct influence on the richness of the reward environment. To further optimize reward income, it is therefore also important to react adaptively to the presence of conspecifics foraging for similar resources. This can happen indirectly via adaptive responses to the changes in the reward environment that others cause, but it can also be that over time in some species computational mechanisms

evolve to predict the consequences others will have on one's own reward rate. This may rest on estimates of competitive ability, which would be in a clear and simple way related to the foraging behavior of the animal, rather than requiring more complex perspective taking. However, in contrast to a mechanism for tracking reward rates over time, basic ability estimation might require more dynamic and flexible representations, as the effects of others on own reward income can change much more quickly than slowly drifting reward rate changes in a patch, particularly if the relationship to others oscillates between cooperation and competition.

In this thesis, I investigated the neural mechanisms supporting the types of behavior outlined above. Chapter 2 and 3 focus on value computations in reward environments with temporally auto-correlated reward rates. Chapter 4 and 5 investigate self and other ability estimation and the representations of social relationships between self and other.

1.2. Methods in cognitive neuroscience

I used functional magnetic resonance imaging and computational modeling to identify neural mechanisms underlying reward tracking and social learning. In the following, I will give a brief overview of these neuroscientific methods as well as related methods that are important in the broader context of this work.

1.2.1. *In vivo electrophysiological recordings*

Electrophysiological recordings, mostly in non-human primates and rats, have long been the most important tool in cognitive neuroscience research. Sometimes microelectrodes are implanted chronically and measure extracellular currents that reflect spiking action potentials of single neurons. Electrophysiological recordings serve as a

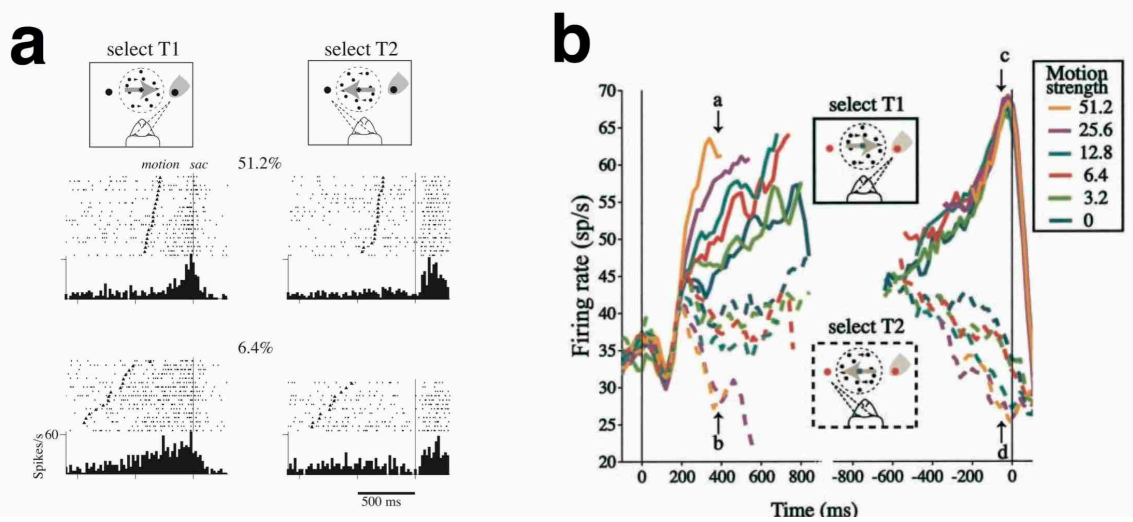


Figure 3. Classic example of electrophysiological recordings (Roitman and Shadlen, 2002). *Results from a study using a random dot motion paradigm in which monkeys saccade to one of two locations to indicate the displayed average motion direction. Performance accuracy and reaction time depend on motion strength, i.e. motion coherence. Neurons are recorded in LIP. Electrophysiological studies usually report responses of single cells (**a**) as well as the whole population of recorded neurons (**b**). Single cell results are shown in raster plots and histograms time-locked to stimulus presentation, or, in this case, response time. The average firing rates in (b) show that cells encode motion strength and responses made. [Panels reproduced from (Roitman and Shadlen, 2002) with permission of Society for Neuroscience; permission conveyed through Copyright Clearance Center, Inc.]*

correlational method to investigate whether cells in a particular brain region encode information that is relevant in the context of the experiment, for example information about stimuli presented or choices made (Fig.3). The information a brain region has access to is a crucial determinant of the computations that the brain region can perform and electrophysiological recordings therefore have important implications for theories about brain function.

A difficulty for studies using electrophysiological recordings is that they are usually limited to one, or a few, recording sites. Often, however, it is of equal importance to test whether other brain regions lack the signals of interest. Neurons with similar properties are frequently found in different brain regions and it is the relative predominance of signals that is informative (Kennerley et al., 2011; Murray et al., 2014; Seo and Lee, 2009; Stoll et al., 2016).

In recent years, analyses of electrophysiological data have become more sophisticated. Rather than calculating simple average firing rates, multivariate approaches analyze higher-dimensional patterns in the firing rates of cell ensembles (Karlsson et al., 2012; Mante et al., 2013; Stokes et al., 2013). This is an interesting and promising development that aims at investigating network dynamics rather than only the coding of information by individual neurons. However, mechanistically, it is often unclear how the multivariate patterns arise from the firing of the individual neurons.

1.2.2. *Lesion and inactivation studies*

Lesioning or inactivating a brain region is another method to study localized brain function. It is used to assess whether activity in a brain region is not only correlated with a variable of interest, but whether a systematic behavioral deficit occurs after lesioning/inactivating it. Hence, it is a causal method that can in principle provide stronger evidence for linking computations to a brain region than correlational methods. However,

with lesions it can sometimes be difficult to individuate and identify the precise computational mechanisms at play in a brain area.

In humans, there are famous case studies of patients with brain damage that provided seminal insights into the role of prefrontal cortex in behavioral control (Harlow, 1868) and medial temporal cortex in memory (Scoville and Milner, 1957). In humans, however, anatomically circumscribed lesions are very rare. Moreover, transcranial magnetic stimulation or transcranial direct current stimulation can be used to (to a very small degree) disrupt localized brain function (Higo et al., 2011; Huang et al., 2005; Polania et al., 2015).

In animals, it is possible to conduct studies with very precise lesions or inactivations to map out functions of anatomical subdivisions, for instance in orbitofrontal cortex of monkeys (Noonan et al., 2010; Walton et al., 2010). It is possible to test for instance whether inactivation of a brain region in which single cell activity correlates with choice behavior leads to choice deficiencies (Katz et al., 2016). A challenge for these types of studies is that neural reorganization or functional compensation can occur that disguise lesion/inactivation effects. Also, it is in practice very difficult to set precise lesions and sometimes lesion effects can occur because of adjacent white matter damage rather than the intended lesion to grey matter (Rudebeck et al., 2013).

1.2.3. *Neuroimaging*

Human functional magnetic resonance imaging (fMRI) has become probably the most frequent method in human cognitive neuroscience. Several software packages are available to analyze task-related brain activity (Penny et al., 2011; Smith et al., 2004). fMRI has the advantage that it can test for brain-wide correlations between neural activity and a task variable or behavioral variable of interest.

fMRI rests on the blood-oxygen-level dependent (BOLD) contrast (Ogawa et al., 1990). These changes in blood oxygenation in a brain region can be assessed with an appropriate MRI sequence. The BOLD signal follows neural activity with a temporal delay, the hemodynamic response function (HRF). The HRF peaks reliably around six seconds after task onset, although there is some variation between subjects and brain regions (Buckner et al., 1996). The BOLD response in a brain region can be linked to local field potentials indicating that fMRI might pick up the input signals rather than the output signals of a brain region (Logothetis et al., 2001). Overall, however, it is not clear which neural network dynamics the BOLD signal reflects and how these dynamics can be linked to the firing rate of single cells. Nonetheless, although it is not entirely clear which conclusions can be made with respect to the underlying network dynamics at the single-neuron level, the BOLD signal can be used to establish links between some sort of neural activity in a brain region and an experimental manipulation or pattern of behavior.

Temporal precision of fMRI is limited by the sluggishness of the BOLD response and delays in the sampling rate of the MRI. Spatial precision is limited by the grid size of the functional images (usually 3mm x 3mm x 3mm), but is more fundamentally limited by the spatially coarse nature of the hemodynamic response and the organization of the vascular system.

Task-fMRI is usually analyzed in an event-related way using a general linear model (GLM). Events of interest need to be temporally or statistically dissociated to find separable effects. Factorial designs are frequent. Here, specific task elements are systematically changed and resulting differences in the BOLD signal between the conditions are related to the cognitive differences between the conditions. Parametric designs, instead, aim to identify how variations in the BOLD response are related to parametric variables of interest. Most fMRI analyses in this thesis are parametric analyses. For example, by systematically varying recent reward rates over trials, it was possible to

identify brain regions where the BOLD signal scaled with the recent reward rates. Parametric fMRI designs make the assumption that the BOLD signal scales linearly with the parameters of interest. However, this assumption can be tested in the data and moreover it is rarely critical for the conclusions whether the relationship is linear or nonlinear (or even a step-function), as long as the regressor is still sensitive for the effect of interest. Parametric fMRI analyses are well suited to be combined with computational models, as these can equip the fMRI analyses with trialwise parametric estimates of cognitive processes of interest.

1.2.4. *Computational modeling*

Computational models in cognitive neuroscience can offer a mechanistic description of cognitive processes underlying human behavior. They can be used to mathematically formulate and compare hypotheses about strategies subjects use in decision-making. They often provide trialwise estimates of subjective decision variables and other latent variables driving choice. Hence, they nicely complement parametric fMRI analyses to identify neural mechanisms of interest.

Throughout this work, I have used computational modeling, in particular reinforcement learning (RL) modeling. The delta rule has had a central role in RL; it provides a simple account of how value learning takes place (Rescorla and Wagner, 1972; Sutton and Barto, 1998). This rule says that a learner forms estimates about the value of currently available options and revises these estimates every time they encounter a new outcome associated with that option. The mismatch between outcome and expectation is the prediction error (PE), which determines, together with the learning rate, how much the current expectation is revised (to the positive or negative). A notable feature of RL models is the unit of time. They operate with discrete time steps that are often mapped on separate trials in the context of psychological experiments. This assumes that the value of an option

is stored and only updated when the next outcome of an option is encountered, unless some sort of forgetting is implemented in the model (Collins and Frank, 2012). There are also many other ways in which models using the delta rule can be extended, for example temporal difference models can account for learning stimulus-reward associations over time (Schultz et al., 1997; Sutton and Barto, 1998), additional information about task-structure can improve model accuracy (Daw et al., 2011), or Bayesian models can incorporate effects of uncertainty into value estimation (Behrens et al., 2007a).

In my investigations of foraging-related value tracking and social behavior, I used simple RL models to approximate the value representations that guide choice in these contexts. RL models served as a simple default model. As useful and fundamental as simple RL models were in these investigations, I found that people's value representations systematically deviated from the predictions of the RL algorithm. First, in ecological environments characterized by auto-correlated and slowly drifting rewards rates, not one simple RL value estimate guides choice; instead, multiple reward memories and their relationships to one other guides choice. Second, in the social domain, value estimates for one self and for other persons were not independent, but instead influenced each other reciprocally depending on the social context.

1.3. Medial frontal cortex: Anatomy

This thesis focuses on a set of regions in medial frontal cortex involved in value computations and the types of foraging-related behaviors outlined above. Figure 4 shows a summary of the regions investigated. They comprise dorsal anterior cingulate cortex (dACC), perigenual anterior cingulate cortex (pgACC), Brodmann area 9, and ventromedial prefrontal cortex (vmPFC).

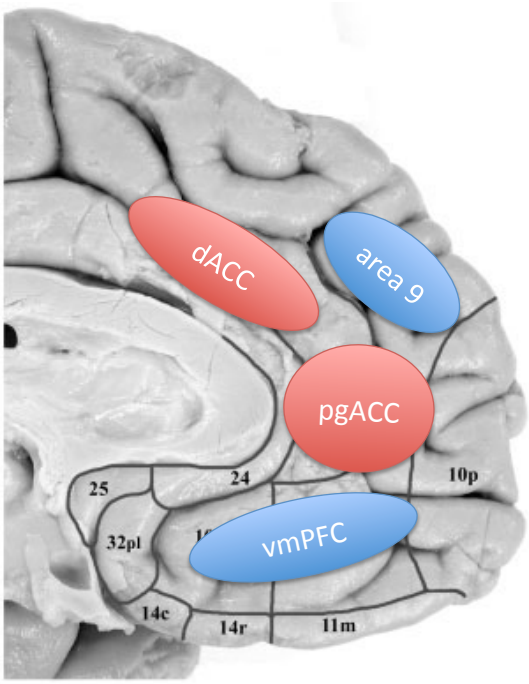


Figure 4. Sagittal view on human medial frontal cortex. *Regions of interest are overlaid. Color coding refers to cytoarchitectonic features of the areas studied. Red indicates agranular and blue indicates granular cortex. Abbreviations: dACC, dorsal anterior cingulate cortex; pgACC, perigenual anterior cingulate cortex; vmPFC, ventromedial prefrontal cortex; area 9, Brodmann area 9. [Reproduced from (Ongur et al., 2003) with permission from John Wiley and Sons.]*

1.3.1. *Dorsal anterior cingulate cortex*

In humans, anterior cingulate cortex comprises the anterior parts of the cingulate gyrus and the cingulate sulcus dorsal to the genu. In addition, in humans, in roughly 50% of people it comprises a paracingulate sulcus and gyrus which is more prominent in the left

compared to the right hemisphere and absent in a hemisphere in *circa* 10% of cases (Paus et al., 1996). This feature can impede precise localization of brain function in these areas across subjects, for instance in the case of mapping out somatotopy in the cingulate (Amiez and Petrides, 2014).

Anterior cingulate cortex can be further subdivided based on cytoarchitectonic features and connectivity. Cytoarchitectonically, frontal areas can be categorized by the presence or absence of granule cells in layer IV. There is a spatial gradient visible with a more prominent granular layer IV in anterior compared to posterior regions of the frontal cortex (Wallis, 2012). Granular cortex is present in humans and monkeys, but not in rats (Wise, 2008). Anterior cingulate cortex is part of the agranular cortex, in contrast to prefrontal areas such as ventromedial prefrontal cortex or dorsomedial area 9. Within anterior cingulate cortex, there is in addition a subdivision between areas 24, 32 and 25, which are present both in humans and macaques and that can even be mapped onto rats' and mice anterior and mid cingulate cortex (Vogt and Paxinos, 2014).

Since the late 1990s activity has been frequently detected in human anterior cingulate cortex in a region dorsal to the corpus callosum in neuroimaging experiments. A somewhat ill-defined region that lies between approximately y=0 and y=30 (Montreal Neurological Institute (MNI) coordinates) in the cingulate sulcus has come to be referred to as the dorsal anterior cingulate cortex (dACC). This functionally defined region approximately corresponds to the anatomical area 24c and a neurophysiologically defined area called the anterior rostral cingulate zone (Picard and Strick, 1996, 2001) (Fig.5). To avoid unnecessary confusion, however, I will refer, when possible, to this region as dACC from here on. A similar region can be identified in both humans and macaques and the way in which its activity is correlated with activity in other brain areas suggests that there are many similarities in the anatomical connections of this region in both humans and macaques (Neubert et al., 2015).

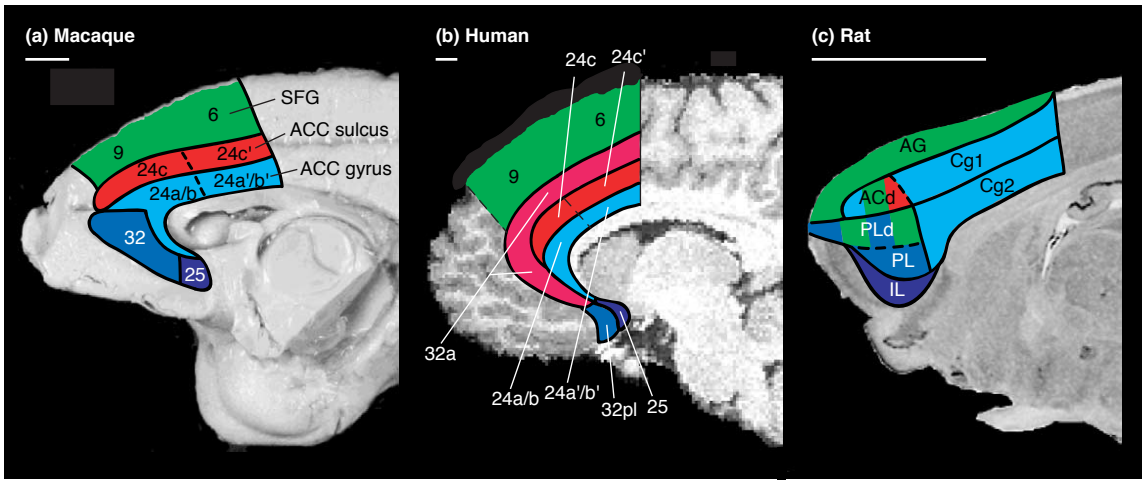


Figure 5. Comparative neuroanatomy of anterior cingulate cortex. *Dorsal anterior cingulate cortex lies towards the anterior end of the anterior cingulate sulcus, in the vicinity of area 24c. As part of the agranular frontal cortex, it is possible to identify similar regions in monkey as well as in rat medial frontal cortex. [Reprinted from (Rushworth et al., 2004) with permission from Elsevier.]*

The connections of dACC reflect its intermediate location between granular prefrontal cortex and areas involved in motor control. In the macaque (Morecraft et al., 2012; Vogt and Pandya, 1987), area 24c is interconnected with prefrontal cortex including its medial-orbital and lateral-orbital surface (areas 10,11,12,13,14), in particular areas 10 and 12. It is densely connected to adjacent subdivisions of the anterior and posterior cingulate cortex and receives inputs from the amygdala and the ventral striatum (albeit to a smaller degree than the more ventral areas 24a,b). In addition, area 24c is connected with dorsolateral prefrontal cortex, in particular area 46 and areas in the vicinity of the temporal pole as well as the lateral parietal lobe. Area 24c has direct connections with the supplementary motor cortex, premotor cortex and motor cortex, which is also reflected in an approximate somatotopic organization in multiple subdivisions of the cingulate in monkeys (Procyk et al., 2014) and humans (Amiez and Petrides, 2014). In a study using resting state functional correlations to infer estimates of connectivity in humans and macaques (Neubert et al., 2015) it was notable that the same region, in contrast to more posterior regions of the cingulate cortex, was more strongly connected to dorsolateral

prefrontal cortex, perigenual anterior cingulate cortex and the frontal operculum at the expense of connections with sensorimotor regions and medial and lateral parietal cortices.

1.3.2. *Perigenual anterior cingulate cortex*

Perigenual anterior cingulate cortex (pgACC) is a medial frontal region anterior and ventral to dACC. It is anterior to the genu and confined by the superior and inferior rostral sulcus and comprises anterior parts of the anterior cingulate gyrus and sulcus as well as parts of the superior frontal gyrus.

Like dACC, pgACC is part of the agranular frontal cortex and as such homologies can be identified in both macaque monkey and in rats (Vogt and Paxinos, 2014; Wise, 2008). In monkeys and humans, pgACC comprises anterior parts of area 24a/b, area 32 and possibly some anterior and ventral parts of 24c (Rushworth et al., 2004; Wise, 2008). PgACC is at the center of the medial wall of the frontal cortex and as such part of the medial network as opposed to the orbital network in frontal cortex (Price, 2007) (note that the medial network also comprises areas at the lateral-posterior edge of the orbital surface). In the macaque, pgACC is strongly connected to surrounding frontal areas on the medial wall but also the lateral orbitofrontal cortex, in particular area 12o, to the cingulate cortex and dorsal temporal regions and very strongly to subcortical areas such as the hippocampus, dorsal hypothalamus and the ventral striatum (Price, 2007). Strong connections exist between pgACC as well as the more ventromedial area 25 and the amygdala, particularly caudal parts of the amygdala (Ghashghaei and Barbas, 2002) (Fig.6). A study examining functional connectivity in humans and monkeys (Neubert et al., 2015) in general confirmed this connectivity pattern in humans, but highlighted strong functional connectivity with dorsolateral prefrontal cortex, precuneus and a respectively stronger coupling of dorsal pgACC parts with the adjacent cingulate and presupplementary

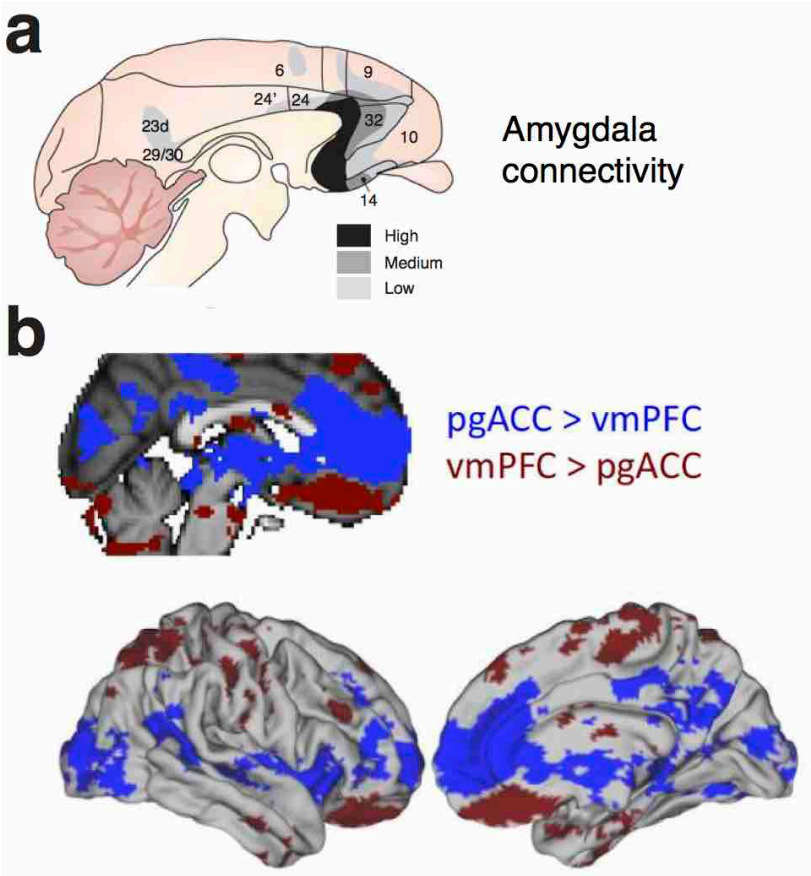


Figure 6. Subcortical connectivity of medial frontal cortex. *Overall, subcortical structures such as the ventral striatum, amygdala and hypothalamus share a similar pattern of anatomical connections to the medial frontal cortex. This pattern is exemplified in panel* ***(a)*** *by showing the projection density from the amygdala to the medial frontal cortex. Note that the densest projections are to regions surrounding the genu of the corpus callosum. Hence, subcortical connections tend to be strongest for dACC, pgACC and area 25 compared to area 9 and some parts vmPFC. [Reprinted by permission from Macmillan Publishers Ltd: (Price and Drevets, 2010)]. (**b**) In line with this, when comparing functional connectivity between pgACC and vmPFC, pgACC tends to cluster together with adjacent dACC and subcortical regions such as the ventral striatum, while vmPFC has stronger connections with very dorsal (pre-)frontal regions such as area 9. [Reproduced from Neubert et al., 2015].*

motor areas and ventral pgACC parts with medial temporal areas and ventrolateral prefrontal cortex.

1.3.3. *Area 9*

Dorsomedial area 9, often referred to as dorsomedial prefrontal cortex (Apps et al., 2016), is also part of the medial wall in prefrontal cortex. It is superior and anterior to the cingulate sulcus in the medial part of the superior frontal gyrus. In both monkeys and

humans, area 9 extends over the dorsal convexity into dorsolateral prefrontal cortex. There is debate which parts of dorsolateral cortex area 9 occupies and how far it extends (Petrides, 2005). In the monkey, cytoarchitectonical signatures of area 9 are still visible in the dorsal bank of the principal sulcus.

In contrast to the previously described dACC and pgACC, area 9 is part of granular prefrontal cortex (Wise, 2008). In monkeys and humans, medial and lateral subdivisions differ in that there is a relatively more developed layer IV (at the expense of layer III) in the medial subdivision and *vice versa* in the lateral subdivision (Petrides, 2005). Area 9 is tightly connected to adjacent granular prefrontal areas (e.g. area 10, area 46) as well as anterior cingulate cortex regions such as area 32 and area 24. Compared to pgACC or dACC, area 9 has more sparse connections to subcortical areas such as the amygdala or ventral striatum, but some connections exist with the midbrain and hypothalamus (Price, 2007). In general, in comparison with pgACC, connections with medial temporal and (peri-)rhinal regions are sparser for area 9. In contrast to other medial granular areas and in particular pgACC, area 9 has considerable connections with medial and mid-dorsal premotor and presupplementary motor areas (Barbas et al., 1999; Neubert et al., 2015; Petrides and Pandya, 1999). Compared to pgACC, area 9 has stronger functional connectivity with ventrolateral and posterior-lateral orbitofrontal areas such as areas 47, area 44v in the inferior frontal sulcus, both in humans and macaque monkeys (Neubert et al., 2014; Price and Drevets, 2010). Note that there is some uncertainty not only on the cytoarchitectonical description of area 9, but also on how to best describe its connectivity profile. For instance, although it is tightly connected to adjacent areas in medial prefrontal cortex, it has been suggested that its function might be better understood as part of a dorsal network in conjunction with adjacent granular areas 46d and the frontal pole (Fig.7) (Price and Drevets, 2010).

In sum, dorsomedial area 9 is tightly connected with adjacent agranular and, in particular, granular frontal regions. It is, in contrast to dACC and pgACC, less connected with motor related regions as well as with limbic medial temporal regions, but has widespread cortico-cortical connections, in particular with the ventrolateral frontal cortex.

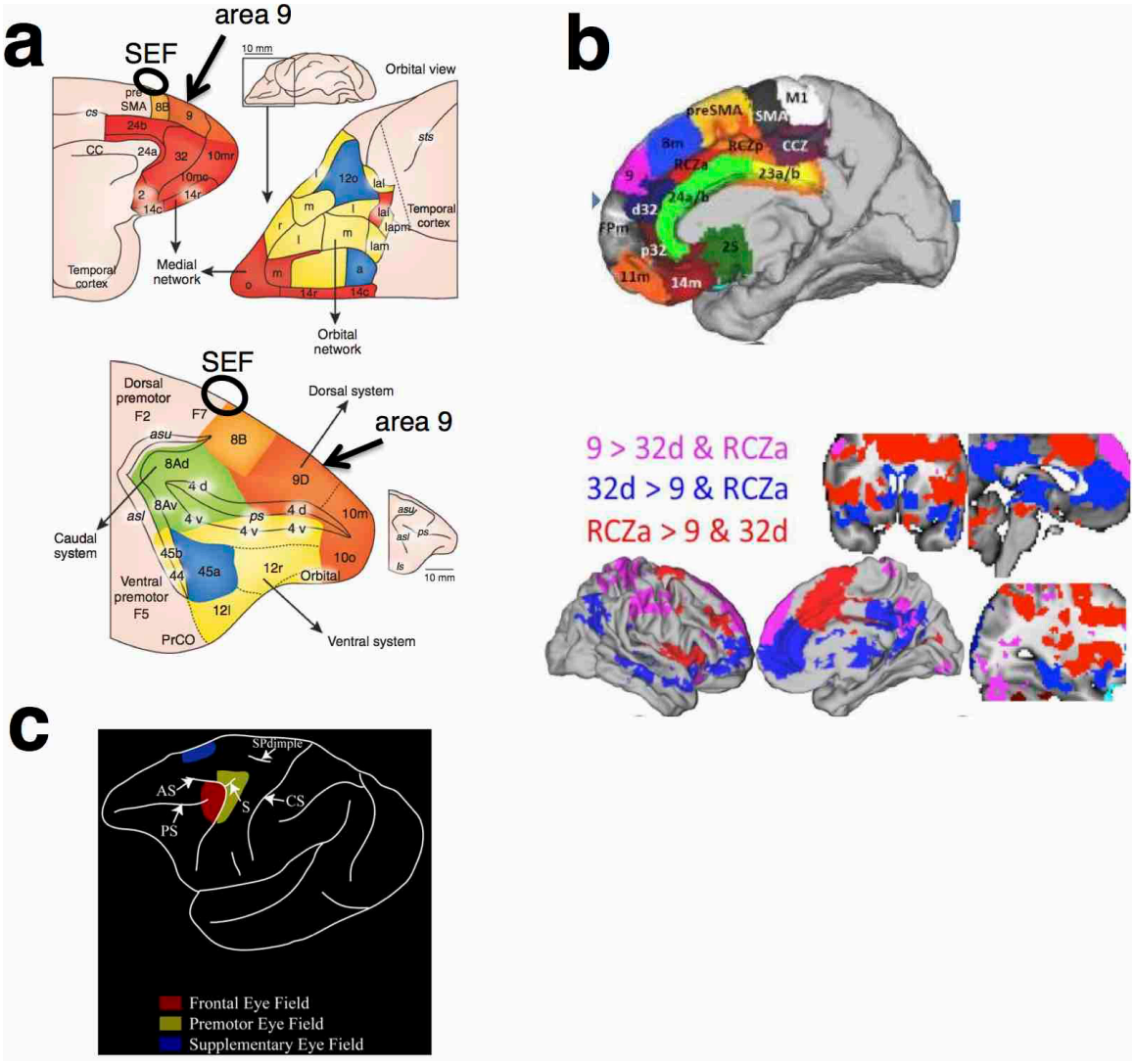


Figure 7. Anatomy of area 9 and supplementary eye fields. ***(a)*** *Sagittal, coronal and transverse illustration of frontal cortex in macaques according to Price and Drevets. (Price and Drevets, 2010). Areas of the same color have strong anatomical connections with each other and similar connections to other brain regions. Note that area 9 (see arrows) has a relatively peripheral position. Circles indicate adjacent supplementary eye fields (SEF). [Reprinted with permission from Macmillan Publishers Ltd: (Price and Drevets, 2010).]* ***(b)*** *DTI-based parcellation of medial frontal cortex in humans. Color coding in lower panel indicates relatively stronger functional connectivity of area 9, area 32d (roughly pgACC) and the anterior rostral cingulate zone (RCZa; dACC) (Neubert et al., 2015).* ***(c)*** *Location of SEF (blue) in macaques. [Reprinted from (Amiez and Petrides, 2009) with permission from Elsevier.]*

It is difficult to identify animal studies that could help to clarify the function of area 9. Area 9 is part of granular prefrontal cortex so it is not possible to identify a homologous area in rats (Wise, 2008). In monkeys, it is possible to identify a homologous region (Neubert et al., 2015; Sallet et al., 2013; Wise, 2008; Yeterian et al., 2012). However, very few studies of function, whether with lesions, inactivations, or neurophysiological recordings have directly focused on area 9. Instead, they often focus on more posterior medial frontal cortex, in particular the supplementary eye fields (SEF), anterior and medial to the frontal eye fields (Fig.7). The SEF is a functionally defined area. Electrical microstimulation in this area elicits eye movement (Schlag and Schlag-Rey, 1987). Unlike the frontal eye fields, SEF stimulation can elicit converging saccades that appear to be goal directed, and saccades appear with a longer latency. The size and exact location of the SEF varies. They are on the dorsomedial surface at the border of area 6 to area 8, dorsal and slightly anterior to the arcuate sulcus (Amiez and Petrides, 2009). It might be that activity in SEF can be taken as a good proxy of computations in area 9, because there are strong connections between the two areas. In particular if decisions are made using eye movements, SEF might provide a good readout of area 9. However, this is obviously only an indirect readout.

1.3.4. *Ventromedial prefrontal cortex*

Ventromedial prefrontal cortex (vmPFC) is another region that has been identified in medial prefrontal cortex on the basis of functional criteria in human neuroimaging studies. It is ventral to pgACC, i.e. ventral and (partly) anterior to the genu, but posterior to the frontal pole. VmPFC is in the vicinity of the (superior and inferior) rostral sulcus.

Similar to area 9, but in contrast to dACC and pgACC, vmPFC is part of granular prefrontal cortex (Wise, 2008). In monkeys and in humans, there has been debate about the exact cytoarchitectonic classification of vmPFC (Rushworth et al., 2011). In monkeys, in

contrast to older classifications that suggest that frontopolar area 10 and caudal area 25 border each other directly (Ongur et al., 2003; Price, 2007), newer classifications suggest the existence of a distinct area 14m in between, just dorsal to the most ventromedial areas 14r and 14c (Mackey and Petrides, 2010). Area 14m contains a very dense granular layer IV. When cytoarchitectonic area 14m is mapped onto MRI scans (Mackey and Petrides, 2014), there is a remarkable resemblance between this area and a corresponding area identified in a DTI-connectivity-based parcellation (Beckmann et al., 2009). Both approaches identify the vmPFC along the rostral sulci and this is in line with the shape of activation clusters found in many functional imaging studies (Basten et al., 2010; Boorman et al., 2009; De Martino et al., 2013a) (Fig.8). This cluster shape is somewhat at odds with a newer parcellation of medial prefrontal cortex (Neubert et al., 2015), which might distinguish more precisely between areas with similar connectivity patterns that are often co-active in imaging studies.

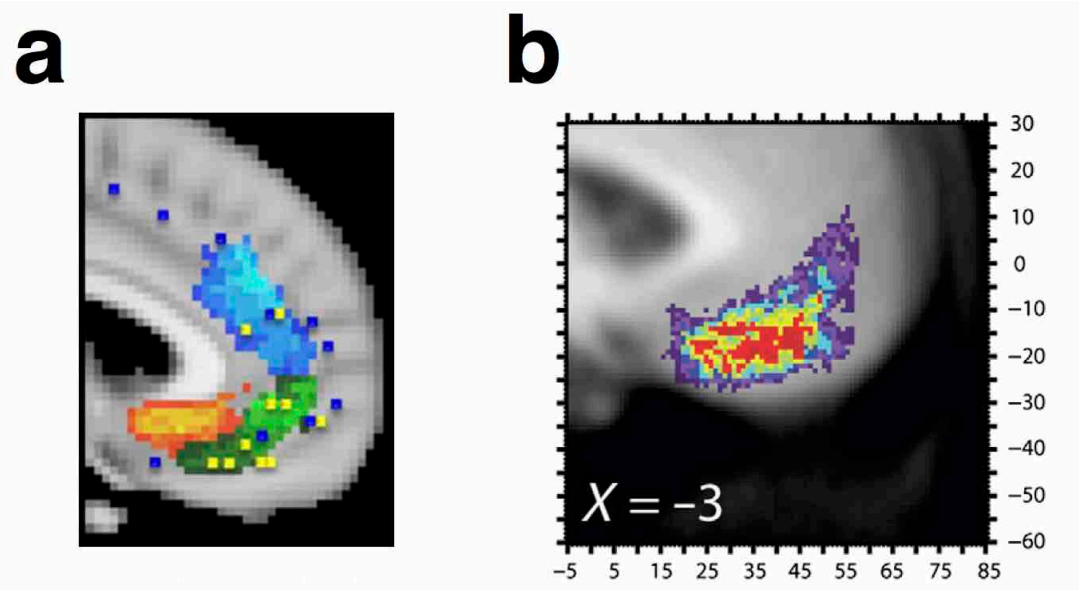


Figure 8. Ventromedial prefrontal cortex. *(**a**) DTI-based parcellation of cingulate cortex identifies a cluster along the rostral sulcus (green cluster). [Reproduced from (Beckmann et al., 2009) with permission of Society for Neuroscience; permission conveyed through Copyright Clearance Center, Inc.] (**b**) A very similar cluster is identified when mapping cytoarchitectonic area 14m on the surface of an MNI brain. [Reproduced from (Mackey and Petrides, 2014) with permission from John Wiley and Sons.]*

Due to the difficulties in cytoarchitectonic mapping of vmPFC, it is not entirely straightforward to establish its precise connectivity profile. It is likely that vmPFC is one of the ventral medial prefrontal areas that lies just at the intersection of the medial and orbital network (Price, 2007) and that would therefore have strong connections not only with other medial prefrontal areas such as pgACC, but also with areas on the orbital surface that receive sensory input from several modalities. Area 14 in the macaque is, as all previously discussed regions, strongly connected to cortical regions that are adjacent to it; in the case of area 14 this encompasses area 25 caudally, area 10 rostrally, areas 32, 24 and 9 dorsally, and areas on the adjacent orbital and lateral orbital surface (Yeterian et al., 2012). VmPFC has only very few, if any, connections with dorsal frontal areas that are involved in motor control, which sets it apart from, for example, area 9, although both areas are part of prefrontal granular cortex (Barbas et al., 1999). VmPFC has, similar to pgACC, strong connections with medial temporal regions like amygdala, the ventral striatum, hippocampus, hypothalamus and rhinal cortices and also with the posterior cingulate and superior and mid-temporal cortex and the temporal poles, although those connections with dorsolateral prefrontal cortex that exist are weaker compared to those of pgACC (Neubert et al., 2015).

1.4. Medial frontal cortex: Function

1.4.1. *Dorsal anterior cingulate cortex*

Dorsal anterior cingulate cortex (dACC) is one of the most-studied and debated brain regions in cognitive neuroscience. Historically, dACC is associated with monitoring behavior for response conflict (Botvinick et al., 1999). This account is based on a simple computational model originally developed for explaining behavior in the Stroop task where incompatible responses are cued by differently colored color words (Cohen et al., 1990). Not unlike more recent accounts of response selection (Cisek, 2012), the model assumes a parallel network architecture. In the network, response tendencies can be active at the same time in a competitive selection process, resulting in response conflict (Fig.9). In its early formulation the theory fits with the neuroanatomy of dACC, as it could be argued that dACC is in an ideal position to monitor and adjust behavioral responses given currently pursued goals.

Over time two related developments made the theory less appealing from a neuroscience perspective. First, the introduction of more general forms of conflict detached the theory from the motor domain and blurred its conceptual precision. For example, it has been suggested that "decision conflict", a generalized type of response conflict which is not bound to specific bodily responses, exists and is linked to a region posterior to dACC at the edge of the supplementary motor cortex (Cole et al., 2009). Second, the theory introduced an explicit processing unit for monitoring conflict (Botvinick et al., 2001) which is at odds with other models of decision making and neurophysiological recordings (Fig.9b). For example, in biophysically plausible models of decision making or drift diffusion type models of decision making (Gold and Shadlen, 2007; Wang, 2008), decision uncertainty exists and can be linked for instance to the duration of the decision process (Hunt et al., 2012) or the difference in subjective value of the available options at the time of choice,

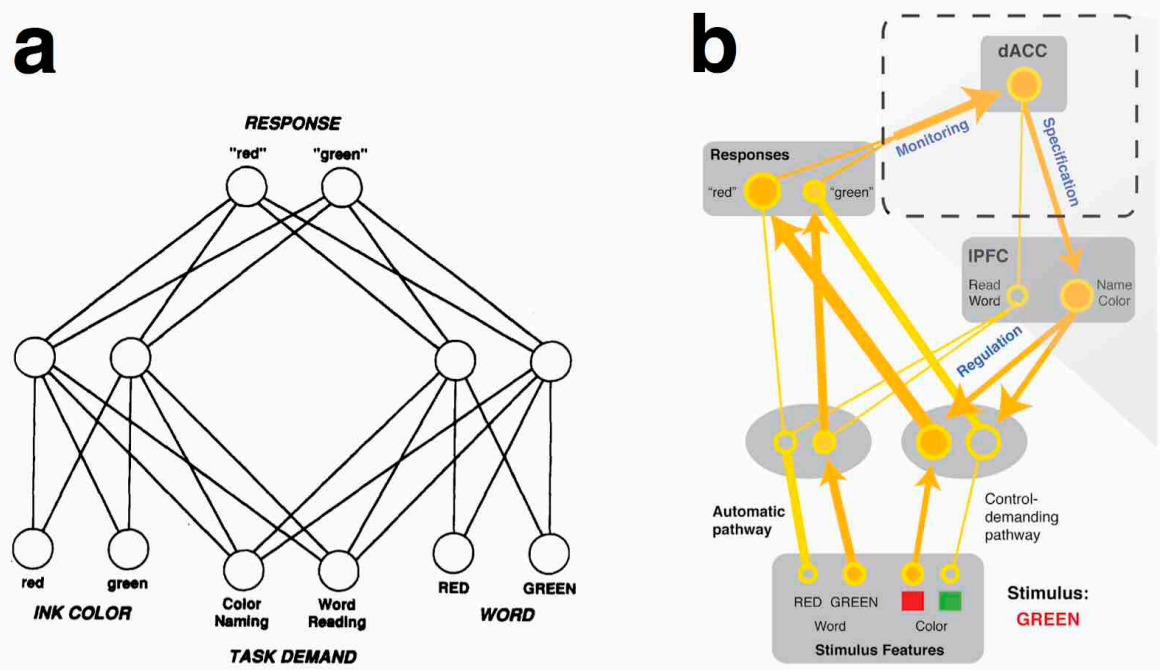


Figure 9. Conflict monitoring account of dACC. *(**a**) Earlier (Cohen et al., 1990) and (**b**) later formulations (Botvinick et al., 2001; Shenhav et al., 2013) of a cognitive control model for the Stroop task. Note that all network nodes are present in both (**a**) and (**b**), except the monitoring unit. It is this unit, and not for instance the response units themselves that are thought to lead to measurable effects in fMRI. (panel a reprinted from (Cohen et al., 1990) with permission, publisher: APA; panel b is reprinted from (Shenhav et al., 2013) with permission from Elsevier, but the critical feature of the theory was already introduced in the 2001 paper from Botvinick and colleagues).*

i.e. decision confidence (De Martino et al., 2013a). Hence, decision uncertainty arising from equally active response units is a common feature of choice models. The critical difference is that the conflict model suggests that a separate functional unit exists in the brain that is dedicated to the detection of conflict rather than conflict being a by-product of for example competing pools of neurons representing action values. On the one hand, this additional feature seems computationally redundant and on the other hand, it is this feature that makes it a considerable problem for the theory that no single neurons have been found that code response difficulty *per se* independently from coding a particular action (Cai and Padoa-Schioppa, 2012; Ebitz and Platt, 2015; Hayden et al., 2011a; Nakamura et al., 2005).

More recent accounts of dACC emphasize its role in behavioral adaption in

conjunction with dorsolateral prefrontal cortex. Quite early findings showed that errors lead to an increase of dACC BOLD activity and post-error slowing on the next trial (Ullsperger and von Cramon, 2001) and that changes in rewards activate dACC neurons when these changes are used to guide behavioral response alternations (Shima and Tanji, 1998). Many subsequent studies showed that dACC is involved in monitoring of choice outcomes, positive or negative, especially when they are used to update response strategies (Amiez et al., 2005; Behrens et al., 2007a; Kennerley et al., 2011; Kennerley et al., 2006; Matsumoto et al., 2007; Nieuwenhuis et al., 2004; O'Reilly et al., 2013). Moreover, reward-related signals in dACC appear to be highly context sensitive. For example, the same neurons in dACC have been shown to code rewards positively or negatively depending on task design (Hayden et al., 2011a; Hayden et al., 2009) and reward coding can depend upon the type of choice that led to the reward (Kolling et al., 2014). It was suggested that this difference relates to the way rewards guide behavior on subsequent trials and that higher dACC activity triggers behavioral change (Hayden et al., 2011a).

The context-dependent format of reward coding in dACC is striking. It raises a question about the frame of reference in which value is represented in dACC. One possibility is that dACC codes rewards relative to previously set goals or tasks. This idea fits with some general features of dACC activity. For example, dACC is active throughout the period that a task is performed and distinct modulations occur at task onset and offset (Dosenbach et al., 2007; Dosenbach et al., 2006). It codes rewards relative to their position in a sequence towards achieving a behavioral goal (Shidara and Richmond, 2002) and exhibits ramping activity while committing to a choice (Blanchard et al., 2015). The high context sensitivity of reward coding has been used to argue that dACC was positioned late in the processing hierarchy, concerned with monitoring choice execution and choice outcomes rather than making choices (Hayden et al., 2011a). Similarly, the lack of evidence of absolute value coding at the time of decision led to the suggestion that dACC

is not involved in decision making per se. For example, Cai and Padoa-Schioppa have investigated economic choice behavior in monkeys and while they can find evidence for chosen-value coding neurons in dACC, they cannot find neurons coding offer values preceding the chosen value representations (Cai and Padoa-Schioppa, 2012). In contrast, such representations have been found in orbitofrontal cortex (Padoa-Schioppa and Assad, 2006). Similarly, Hayden and colleagues argue that representations of chosen and unchosen value in dACC are not sufficient to implicate dACC in decision making, but that signatures of preceding absolute value coding is necessary as well (Blanchard and Hayden, 2014). However the existence of chosen and unchosen value signals was precisely the evidence that was first used to argue that vmPFC is concerned with decision making in human imaging studies (Boorman et al., 2009; Hunt et al., 2012).

The question whether dACC is involved in decision-making is a difficult one. The wealth of value representations in dACC and their clear relationship to action execution make a role in decision-making very plausible. What is difficult is that it is not entirely clear what would qualify as a proper "decision". If one employs a specific, very restricted notion of "economic decision making" (Cai and Padoa-Schioppa, 2012), then it seems fair to say that OFC plays a more central role in such decisions compared to dACC. However, these types of decisions are usually one-shot events immediately followed by reward receipt; with this narrow notion of decision, there is usually no need to monitor decisions or goal pursuit over a longer period of time. In contrast, there is evidence for dACC's involvement in behavioral adaptation and when employing a broader notion of decision making, the behavioral changes triggered by dACC may well qualify as decisions (Rushworth et al., 2012). From this perspective, dACC might be involved in decisions that take place preferentially (but not exclusively) at the time of feedback when positive or negative outcomes are witnessed and response strategies for subsequent time points are evaluated. This type of decision maps well on to the leave-stay decisions that animals have

when foraging, as discussed in previous sections. While economic decisions are made over simultaneously available options, here the choice is usually to engage with a current choice or to search for alternatives. Not always (Kolling et al., 2012), but often, this choice is embedded in a temporal sequence where the decision is to continue with the current behavior or to change and to switch to an alternative. Note that from this perspective decisions are very similar to what happens during behavioral adaptation when feedback is received: a choice pattern is maintained or altered.

Such an account is compatible with observations of contextual value coding in dACC, as one would expect rewards to be represented relative to a current or default choice. Notably, this account fits with observations not only of feedback-related activity in dACC, but it also predicts activity at the time of choice if the decision has the appropriate format. Hayden and colleagues have found that firing rates in dACC increase up to a threshold when making the decision to stay in a current environment or to leave it for a different environment (Hayden et al., 2011b). Similarly, in humans, signals in dACC scale with the value of engaging with a current option and also (but inversely) with the value of abandoning it (Kolling et al., 2012). In the latter study, it is in particular noteworthy that both signals were present irrespective of the choice taken - providing evidence for the type of absolute value coding that some people argue is required in a decision making region (Blanchard and Hayden, 2014; Cai and Padoa-Schioppa, 2012).

It seems clear that dACC holds a multitude of value signals. It computes the values of actions embedded in a longer-time task context or environment (Kolling et al., 2014; Shidara and Richmond, 2002) and rapidly updates them based on new information (Nieuwenhuis et al., 2004; O'Reilly et al., 2013). What are the computational requirements for such a rich and flexible representation of the reward history? Lee and Seo and colleagues have, over many years, systematically investigated how not only current but also previous rewards influence single cell firing rates in dACC in macaques (Fig.10).

They mostly used versions of a "matching pennies" task in which monkeys played a competitive game with a computer competitor. In this binary choice task, the monkey makes choices by making a saccade to one of two points on a computer screen and receives reward if the computer makes the same choice. The computer is programmed to outplay the monkey. This is done by exploiting patterns in the monkey's previous choice behavior, which in turn incentivizes monkeys to chose as randomly as possible ignoring the previous history of choices and outcomes. Monkeys learn to perform well on this task. However, monkeys' choice patterns are still characterized by traces of a reinforcement learning like the win-stay-lose-shift strategy. Monkeys repeat choices that have led to reward and switch away from choices that have led to losses or no reward (Seo and Lee, 2007). In dACC, firing rates in a trial not only distinguish between rewarded and unrewarded trials, but they also, to a lesser degree, carry information whether rewards were received on previous trials (Seo and Lee, 2007). This could indicate, in line with the RL-like behavioral strategy, that the firing rates encode a longer-term estimate of value, to which current and past rewards contribute.

In a subsequent study (Bernacchia et al., 2011), the contributions of recent and past rewards to neural activity in dACC was analyzed even more rigorously. The authors tried to explain the firing rates of single neurons using the past reward history in a window of 5 trials, taking also into account the current phase within a trial (neural activity was massively influenced by trial phase). They argued that past rewards could affect firing rates in different ways. First they suggest that they might be influenced in a manner consistent with a simple exponential function where the weights of influence decrease over time, second, a double exponential function, or that firing rate is not affected by past rewards at all (Fig.10c). The simple exponential function corresponds to the way a simple RL mechanism integrates rewards to a value estimate (Sutton and Barto, 1998), whereas the double exponential shape is consistent with the way prediction errors weight recent

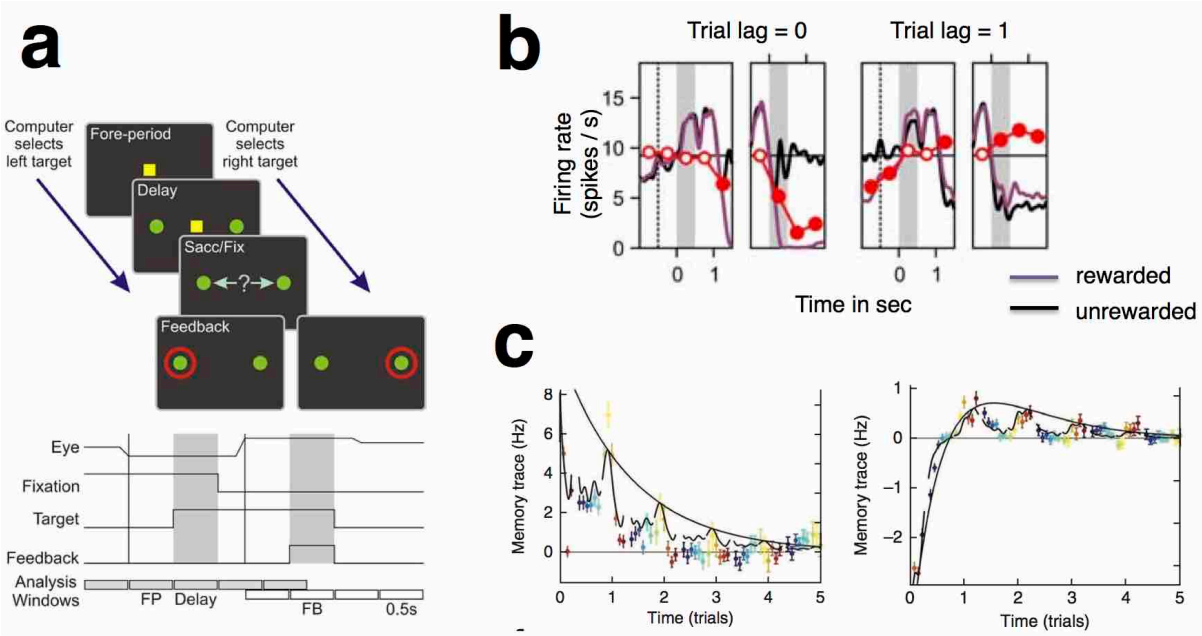


Figure 10. Matching pennies task and single cell firing rates in dACC. *(**a**) Schematic of a version of the matching pennies task used by Lee and colleagues. On each trial, the monkey makes a saccade to one of two targets after a delay cue disappears. The monkey is rewarded if the computer chooses the same target (computer choice indicated by red circle). The grey bars at the bottom match with the grey bars in panel b indicating trial phases. (Seo and Lee, 2007) (**b**) Firing rates from a dACC neuron in response to rewarded (purple) and unrewarded trials (black). In the left panel trials are sorted by whether the current trial was rewarded/unrewarded (Trial lag = 0). In the right panel trials are sorted by whether the previous trial was rewarded/unrewarded (Trial lag = 1); current trial reward effects are averaged out. Red circles indicate difference between rewarded and unrewarded trials; filled circles indicate significance. The example neuron is sensitive to instantaneous rewards but its firing rates are also one trial later still influenced by reward receipt in the previous trial (from Seo and Lee, 2007). (**c**) Two example neurons showing an exponential (left) and double-exponential (right) decay function of memory traces. Discs of same color indicate trial phases, which massively influence firing rates (see panel b). Note the black curves describing the memory function. [panels a and b are reproduced from (Seo and Lee, 2007) with permission of Society for Neuroscience; permission conveyed through Copyright Clearance Center, Inc.; panel c is reprinted by permission from Macmillan Publishers Ltd: (Bernacchia et al., 2011).]*

rewards (Bayer and Glimcher, 2005). The authors find a broad distribution of τ values, ranging from milliseconds to several seconds. The τ parameter characterizing the exponentials corresponds to the learning rate of an RL model. This strongly suggests that neurons in dACC integrate rewards over multiple time scales; some represent recency-weighted reward estimates, whereas others integrate rewards over many seconds.

The Bernacchia study leaves some questions open. For example, when fitting

reward memory exponentials to the neurons, the authors test single and double exponential fits against the baseline assumption that a neuron has no reward memory at all. Given that it is well known that dACC codes immediate rewards and reward expectation, a fairer baseline assumption would be that neurons are only affected by a single past reward event or that they all have a single time constant and then to test whether models with multiple reward constants yield a better fit. It is also of note that no other filter functions were tested. This makes the double exponential the only type of function that allows rewards at different time points to have both negative and positive effects, respectively, on the firing rates. Moreover, in dACC around 70% of neurons had time scales shorter than one trial (one trial took roughly four seconds). So, although there were relationships between the time constants and behavior, it seems unlikely that many of the neurons reflected precisely the type of value estimates that are used in reinforcement learning to compute a recency-weighted average over many trials; note that with this task monkeys usually have rather slow learning rates (Seo and Lee, 2007, 2009). Moreover, reward memories discovered in the study were not analyzed with respect to their relationship to choice in the sense that they were analyzed irrespective of which choices led to reward. In this sense, the study provides evidence that the general rate of reward in an environment is computed with multiple time constants, rather than indicating that the value of specific choices is computed with multiple time constants.

Computing rewards with multiple time constants is not a unique feature of dACC. A similar reservoir of reward time constants was discovered in dorsolateral prefrontal cortex (DLPFC) and lateral intraparietal cortex (LIP) (Bernacchia et al., 2011), with the longest time constants (i.e. longest reward memories) being in dACC. Across the three brain regions, there is a good correlation between the overall reward time constant and the intrinsic time constants of a brain region, i.e. the rate by which random, non-reward related information decays in neurons (Murray et al., 2014). Effects of previous rewards (in

addition to current rewards) on neuron's firing rates have also been found in other brain regions such as dorsomedial frontal cortex near the supplementary eye fields (Donahue et al., 2013; Seo and Lee, 2009) and orbitofrontal cortex (Kennerley et al., 2011; Padoa-Schioppa, 2009).

In sum, signals in dACC dynamically reflect the value of feedback and choice options given the ongoing behavior of the animal. DACC represents rewards with multiple time constants. Although such patterns are also a feature of other brain regions, dACC might be particularly suited to exploit this richness of reward history representations to guide behavior when maneuvering through natural environments with slowly drifting reward rates.

1.4.2. *Area 9*

Compared to, for example dACC and vmPFC, area 9 has received relatively little attention in neuroscience research. Perhaps most prominently, area 9 activity is found in fMRI studies investigating social cognition (Amodio and Frith, 2006), often in conjunction with temporo-parietal junction (TPJ) cortex and precuneus (Saxe, 2006). Classically, area 9 is active when making inferences about the beliefs of other people such as in the verbal false belief task (Fletcher et al., 1995) or when playing competitive games with other people (Hampton et al., 2008; Rilling et al., 2004). It is active when making judgments about another person's traits (Denny et al., 2012), particularly for dissimilar other persons (Mitchell et al., 2006). More recently, this region has been implicated in comparing another person's and one's own preferences (Izuma and Adolphs, 2013). Signals relating to one's own preferences and to dissimilar other's preferences can be found there, if these preferences are not guiding current choices (Nicolle et al., 2012). Moreover, when one person is watching and predicting what another person will do then the difference between the observer's expectation about what the actor will do and what the actor actually does, a

type of prediction error, can be seen in this region (Suzuki et al., 2012; Suzuki et al., 2016). In sum, several other-related signals concerning the predicted actions or intentions or preferences of other people can be found in area 9, but we have only a very limited picture of the type of social- and other-directed computations that are performed in area 9.

Functional accounts of area 9 suggest that this area is involved in mentalizing about the intentions and mental states of other people (Amodio and Frith, 2006) and that this might be performed by simulating another person's intentions and beliefs based on how we ourselves would feel or act in a given situation (Mitchell, 2009). However, note that these suggestions often do not distinguish between subdivisions of medial prefrontal cortex and are thus often not specific for area 9 (Buckner and Carroll, 2007). A conceptually and anatomically more specific suggestion is that area 9 might be involved in triadic attention, in which two agents (not necessarily but often including the self as one of the three) pay attention to the same object or task (Saxe, 2006). There is very little evidence for this suggestion (Walter et al., 2004), but this account is compatible with representations of others as well as self in area 9 and is consistent with area 9's involvement during interactive games. Broadly speaking, it would mean that area 9 computes the functional relationships between agents with respect to some sort of object or goal. However, it is unclear which neural mechanisms would be necessary to achieve this.

As pointed out when discussing area 9's anatomy, there are only a few animal studies than can help us understand the function of area 9. Grey matter in a border region between areas 9, 10, and 46 is increased in macaques that live in bigger social groups and in those that have a higher relative social dominance status (Noonan et al., 2014; Sallet et al., 2011). This indicates that area 9 is also related to social cognition in monkeys and suggests that computations performed in area 9 are important for successful social interactions.

More recently, there has been a series of studies, again from Lee, and Seo and

colleagues recording single cell activity in dorsomedial frontal cortex. Again, they used versions of the matching pennies task (see dACC section). They recorded activity from what they label as dorsomedial prefrontal cortex (Seo et al., 2014; Seo and Lee, 2009) and supplementary eye fields (Donahue et al., 2013). On close inspection, the recording sites for these studies appear to be very similar with the majority of neurons being recorded at the border of premotor area 6 and area 8 (Fig.11). As discussed in the anatomy section, data from these adjacent sites might still give some insight into the function of area 9. I will refer to these regions collectively as dorsomedial frontal cortex (dMFC).

Similar to dACC neurons, dMFC neurons responded to instantaneous rewards, but also to rewards from previous trials (Seo and Lee, 2009). However, in contrast to dACC, dMFC neurons were also modulated by which choice was made, not only on current but also on previous trials. Choice in this experiment was coded as either a leftward or rightward saccade. So there were neurons that, for example, fired more if the monkey had made a rightward saccade compared to a leftward saccade on the last trial, independent of the choice on the current trial. Moreover, some neurons were sensitive to the interaction of choice and reward. Such choice and choice-reward interaction neurons were much fewer in dlPFC and almost absent in dACC (Donahue et al., 2013; Seo et al., 2014; Seo and Lee, 2007, 2009). DMPC therefore carried fine-grained information not only about the current and past rate of reward in the task, but also about the history of choices and which choices had led to reward and which choices had not. So although the monkeys in the matching pennies task should ideally ignore all this information and perform randomly from trial to trial, the monkeys' behavior was influenced by exactly these pieces of information.

As a side note, it is remarkable that past choices could be decoded from dMFC or dlPFC but not from dACC. This might, however, partly relate to the nature of the choices. In a recent study, left/right choices were also found to be decodable from dlPFC but not dACC activity (Stoll et al., 2016). By contrast, however, decisions to maintain

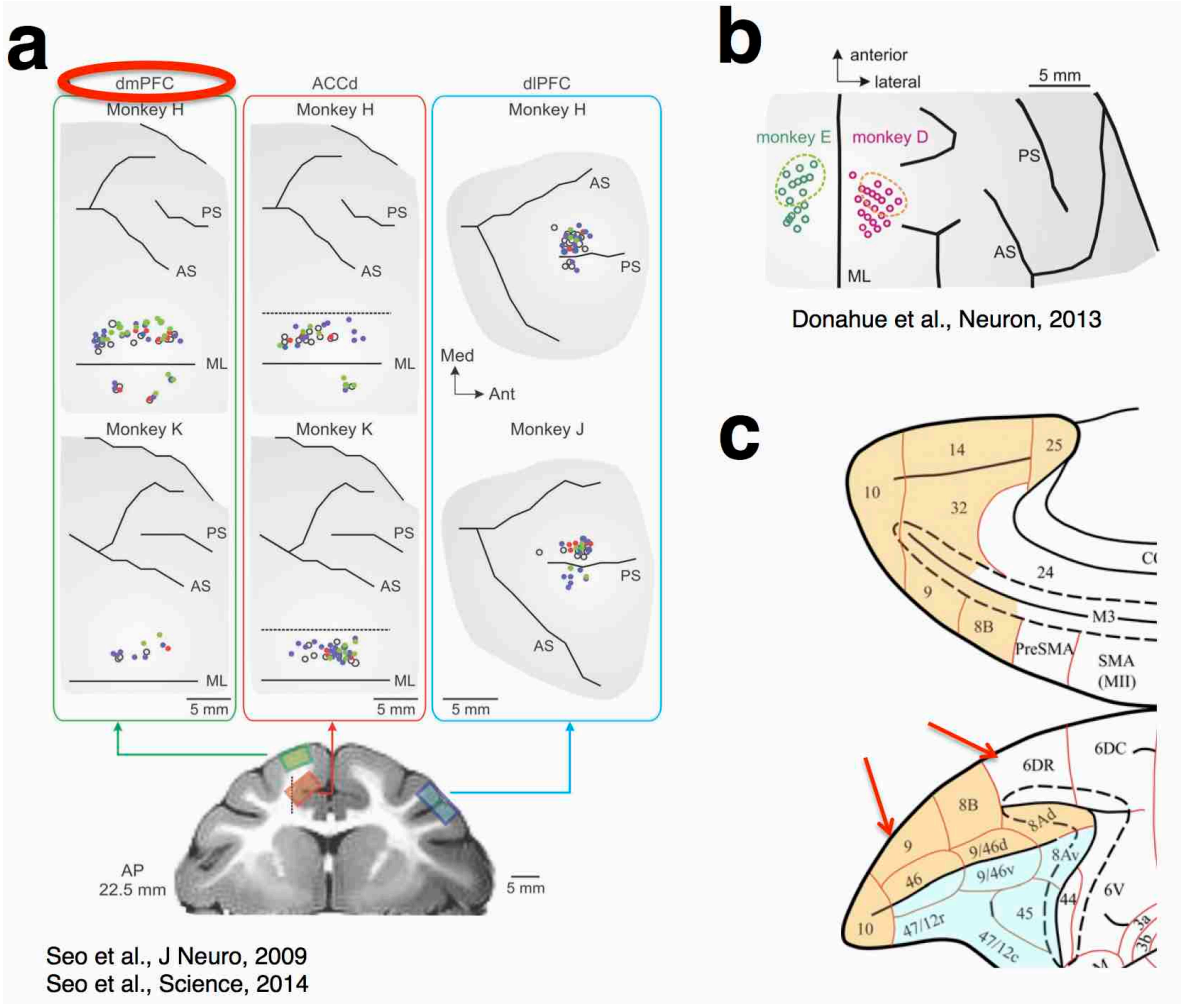


Figure 11. Area 9 related recording sites from the Lee lab. *(**a**) Illustration of recorded cells from "dorsomedial prefrontal frontal cortex" (left column) in two studies from Seo and colleagues (Seo et al., 2014; Seo and Lee, 2009). The color coding refers to reward sensitivity of the recorded cells. [From (Seo et al., 2014). Reprinted with permission from AAAS.] (**b**) Illustration of recorded cells from the supplementary eye fields (dashed lines) from Donahue and colleagues (Donahue et al., 2013). Note that the monkeys from this study did not participate in previous studies from the lab. [Reprinted from (Donahue et al., 2013) with permission from Elsevier.] (**c**) Cytoarchitectonic map of the macaque brain. Note that area 9 (orange; red arrow) is anterior to areas 8 and 6, where most of the recordings from the Lee studies seem to come from. [Reprinted from (Yeterian et al., 2012) with permission from Elsevier.]*

behavior or explore another sub-part of the task to acquire information about reward proximity were better decoded from dACC than dlPFC. While dlPFC may be encoding choices in a more intuitive left-right framework, dACC may be encoding choices in a different framework that might map better onto a distinction between exploration and exploitation. It would be interesting to know whether dACC might carry a memory of

previous decisions if the decisions could be distinguished along different criteria more similar to those used by Stoll and colleagues.

The joint neural representation of past choices and past rewards was closely related to current choice behavior in the matching pennies task. The modulation of neurons' responses to past choices by whether they had been rewarded is predictive of choice switching (Donahue et al., 2013). Seo and colleagues subsequently analyzed the complexity of the dMFC activity patterns in even more detail (Seo et al., 2014). At a behavioral level, they found that monkeys were better able to deviate from an RL-type choice strategy after specific conjunctions of choices and outcomes. As already mentioned, deviation from an RL choice strategy is adaptive in these tasks because the computer program running the task was designed to exploit any RL-based regularities in the monkeys' response patterns. The choice-outcome conjunctions preceding adaptive and maladaptive behavior were calculated over just two trials. Neurons in dMFC had very specific firing profiles for each conjunction and it turned out that the firing rates could be best explained as signaling the upcoming choice that would counteract the RL-based choice pattern. Hence, the rich representation of reward and choice information in dMFC can generate non-RL-based choice. What is not quite clear from these studies is whether dMFC represents others' actions. This is because an unusual feature of the matching pennies task is that rewards are a deterministic function of the monkey's and the computer's choice (if monkey and make the same choice, then the reward will be one point, etc.); from knowing the monkey's choices and rewards, the choice pattern of the computer opponent can be perfectly reconstructed. Hence, it is unclear if dMFC in these studies carried a separate representation of the opponent's choices, as opposed to the monkey's own choices and the reward payoff. In contrast, this also highlights that in the context of social competitive games, the key computation necessary to perform well might relate to being able to use choices and outcomes in a complex, non-RL-based way. One could speculate

that a general feature of social cognition is that it is different from RL-based choice (Lee and Seo, 2016; Seo et al., 2014) and this idea might fit particularly well in the context of competitive games and (higher-level) strategizing. However, an open question remains as to what exactly those non-RL based features of social cognition are and whether they are employed in other domains of social cognition or even non-social cognition.

The specificity with which neurons in dMFC track past choice-reward conjunctions is striking. However, to compute choice values adaptively in different contexts, dMFC should be able to change its coding scheme flexibly. Indeed, the reward-related modulation of neural responses to past choices was specific to the matching pennies task and was not found in a visual search task, where no across-trial contingencies existed (Donahue et al., 2013). Similar results were found in a study examining single cell activity in dorsomedial frontal cortex area 8B, probably slightly anterior (and not overlapping with) the supplementary eye fields (Matsuzaka et al., 2012). The study used a Stroop-like task where colors cued left or right response button presses that could be concordant (color location matches response location) or discordant (color location does not match response location). A few cells fired in a pre-movement initiation period and seemed to be coding for the conjunction of the current rule (concordant/discordant) and the current choice (left/right). In control tasks containing either only concordant or only discordant rules, dMFC neurons turned task-silent.

In summary, area 9 activity is often found in studies of social cognition. In humans, area 9 tends to be more involved in other-related compared to self-related processing, but signals found in this area are not restricted to other-related signals. Animal studies conducted so far are of limited help in refining our understanding of this brain area because they rarely investigated area 9 directly. However, studies from adjacent posterior dMFC show remarkably specific memory traces for past actions and rewards. The neural computations can rapidly change depending on task and can systematically diverge from

reinforcement-guided value formation. This potentially enables higher-order strategizing by detaching response selection from the immediate reinforcement history. This might underlie a special role for area 9 in social interaction where it can be beneficial to be unpredictable. A special role for area 9 in social interactions might also be consistent with its connectional anatomy: area 9's relative proximity to motor-related brain regions and its relative distance to subcortical reward regions. However, the exact neural and computational mechanisms behind area 9's involvement in social cognition remain largely unclear. In the domain of strategic games, unpredictable and deliberate non-RL based behavior might be desirable and mimic the behavior of very clever human poker players, but in everyday life predicting self and others' behavior successfully is often simple and effortless.

1.4.3. *Perigenual anterior cingulate cortex*

PgACC is an anatomical label rarely used in human imaging studies. Although clearly different in terms of cytoarchitecture and connectivity from other prefrontal subdivisions such as vmPFC, dACC and area 9, pgACC activity is often reported as being part of dACC or vmPFC. Perhaps partly for this reason a well-defined model of pgACC function that distinguishes it from adjacent brain areas has yet to emerge and hence, there is also no dominant theme that guides research on pgACC in the same way that notions about "behavioral adaptation", "decision making", or "mentalizing" have recently guided research into dACC, vmPFC, and area 9.

PgACC activity is found in studies of reward-guided choice in humans. Using fMRI, the subjective value of delay discounted rewards correlates with the BOLD signal in pgACC (Kable and Glimcher, 2007). In this experiment, subjects have to evaluate a single option, composed of a monetary reward and a time delay (1-180 days), against a stable reference offer of receiving $20 immediately. The subjective value explained more

variance in pgACC than the component parts of the delayed reward offer (Fig.12a). This means that the BOLD signal in pgACC varies with the individual differences in the discounting rate, i.e. with the individual predisposition to trade-off time against reward. PgACC was again identified in a follow-up experiment where subjects either chose to monitor a single option over time until reward appeared, or to disengage if the reward failed to appear (McGuire and Kable, 2015). It was also found in other follow-up studies involving the evaluation of a delayed reward in comparison to a stable, immediate reference sum (Rodriguez et al., 2015), but in conjunction with vmPFC when both available option changed over time and had a reward and a delay component (Nicolle et al., 2012). Note in some of these studies, pgACC activation might extend anteriorly into area 10 or vmPFC.

It is unclear which feature of delayed discounting tasks leads to the preferential engagement of pgACC. For instance, it might relate to the temporal or abstract nature of the delayed option, its inherent combination of reward and cost, or the fact that the delayed options are judged against a stable reference value. The last point is tentatively consistent with a recent suggestion that pgACC is rather more concerned with evaluation of the multiple factors that might determine even a single option's value than with the process of value comparison between one option and another (Grabenhorst and Rolls, 2011). Such a "direct" role for pgACC in valuation would be consistent with findings of supramodal pgACC activity in emotion recognition tasks (Peelen et al., 2010; Skerry and Saxe, 2014). VmPFC involvement, in contrast, is often found when two concrete offers are compared (Hare et al., 2011b). However, this is not exclusively the case. VmPFC is also active when evaluating a single specific choice only (Hare et al., 2008; Plassmann et al., 2007) or when making choices between options that require integration of a reward and cost components such as reward magnitude and reward probability (Boorman et al., 2009). So, it is difficult to find a clear functional distinction between pgACC and vmPFC.

Alternatively, one might speculate that given pgACC's anatomical connections with both vmPFC and dACC, pgACC might, in general, integrate reward goals with the costs of the action that is required for obtaining the reward. However, usually studies about trade-offs between rewards with physically effortful action costs identify striatal areas and more posterior anterior cingulate regions and pre-SMA rather than pgACC (Bonnelle et al., 2016; Croxson et al., 2009; Scholl et al., 2015).

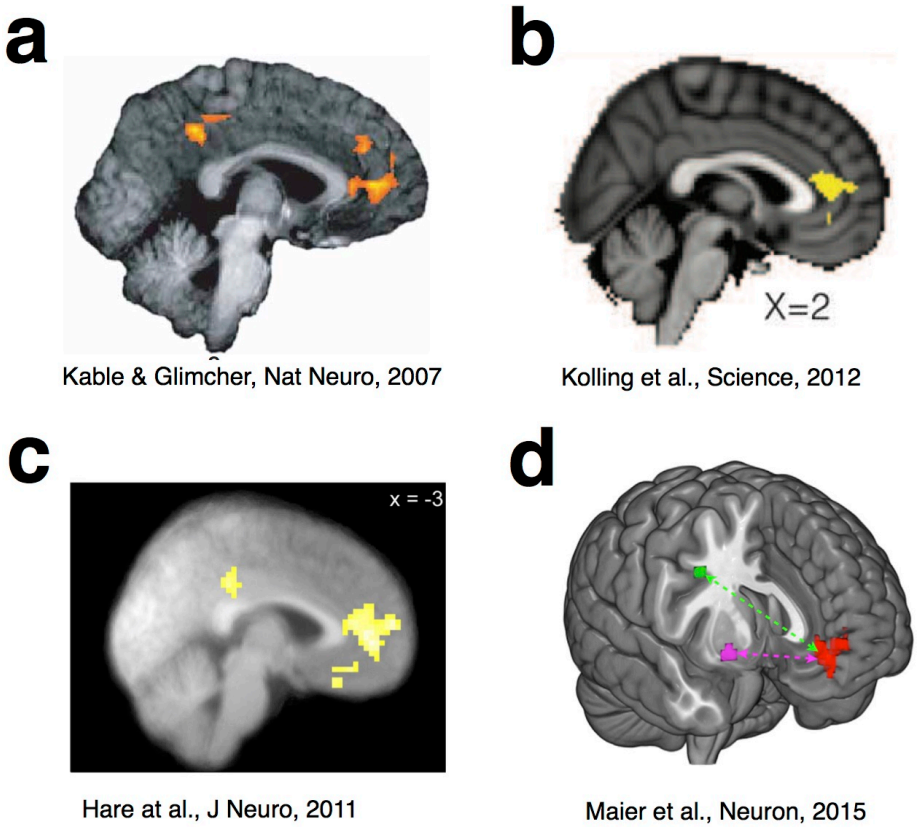


Figure 12. PgACC activity in studies of reward-guided decision making. *PgACC is often active when considering a single option involving a cost-benefit trade-off. Subjective preferences in this trade-off can be read out from this region.* ***(a)*** *Subjective value representation in pgACC in a delay discounting experiment.[Reprinted by permission from Macmillan Publishers Ltd: (Kable and Glimcher, 2007).]* ***(b)*** *PgACC codes the individual bias to make an exploration choice, independent of relevant decision variables, in a foraging task. [From (Kolling et al., 2012). Reprinted with permission from AAAS.]* ***(c)*** *Subjective value of a food item entailing the trade-off of tastiness and healthiness is represented in pgACC. [Reproduced from (Hare et al., 2011a) with permission of Society for Neuroscience; permission conveyed through Copyright Clearance Center, Inc.]* ***(d)*** *The same variable, again represented in pgACC. Purple and green lines indicate individual changes in functional connectivity to the ventral striatum and dorsolateral prefrontal cortex. [Reprinted from (Maier et al., 2015) with permission from Elsevier.]*

The type of accept/reject decisions used in these delayed discounting tasks is also vaguely reminiscent of ideas of dACC's role in engage/forage or stay/switch choices (Rushworth et al., 2012). In a recent foraging study (Kolling et al., 2012), pgACC activity correlated with the general tendency of subjects to make an exploration choice and it was related to the costs that an exploration decision entailed. Note that pgACC was only indirectly related to the actual choice process in that experiment (Fig.12b). PgACC did not track most of the decision variables in that study, neither with respect to foraging decisions nor economic decisions. However, similar to the delay discounting findings, individual predispositions to trade-off rewards for costs were associated with individual differences in pgACC activity.

The suspicion that pgACC might be related to very subjective attitudes towards cost-benefit computations is partly confirmed by studies of self-control (Clithero and Rangel, 2014; Hare et al., 2009; Hare et al., 2014; Hare et al., 2011a). In these studies, subjects have to consider food items that vary along the dimensions of tastiness (reward component) and healthiness (cost component). Subjects rated the overall desirability of the food items. Sometimes they were also cued to attend a single dimension. As in the delay-discounting tasks, a single option was usually considered at a time, possibly in relation to a pre-defined default option. The subjective value of the food items (i.e. after integrating reward and costs) is represented in medial prefrontal cortex (Fig.12c,d). The activation clusters can vary considerably between studies; sometimes they are centered on subgenual vmPFC (Hakimi and Hare, 2015) and sometimes on pgACC (Hare et al., 2011a) or there are separable activations in both areas (Hare et al., 2009). It has been suggested that the relative weights of the cost-benefit evaluations in pgACC are influenced by functional interactions with subcortical structures and dorsolateral prefrontal cortex (Maier et al., 2015). Such a view of the functional connections determining pgACC's activity would fit

with the anatomical connectivity fingerprint of pgACC, arguably even better than for vmPFC.

In macaque monkeys, pgACC is also implicated in cost-benefit decisions (Amemori and Graybiel, 2012). The monkeys performed an approach-avoidance task, in which they made decisions to accept or reject an offer that comprised a variable juice-reward component and a variable air puff-cost component. Cost-benefit trade-off functions were calculated expressing the size of the costs the animals were willing to endure to gain rewards. Microstimulation in pgACC, but not elsewhere, changed the cost-benefit trade-offs: animals were willing to endure fewer costs to receive the same rewards. No similar change was observed in a binary choice task with two rewarded choices suggesting that pgACC manipulation has a highly specific and causal for the relative weighting of costs and benefits rather than a general effect on all aspects of decision-making. Moreover, although the cost-benefit trade-offs were altered by pgACC stimulation, the overall dynamic of choice taking was still preserved. Instead it was just the individual's preferences that were altered.

In summary, in studies on reward-guided behavior, pgACC is implicated in cost-benefit decision-making. Its role is difficult to dissociate from vmPFC's role as fMRI activation clusters for the two areas are often closely adjacent and sometimes overlapping. However, there is some evidence that pgACC activity reflects choice values particularly in situations when there is an accept/reject decision focusing on a single option. Rather than comprising and weighing a wealth of decision variables simultaneously, pgACC's role in choice seems to be subtler. Individual variation in pgACC's activity is related to individual variation to trade-off delay against reward (Kable and Glimcher, 2007), the general willingness to accept costs for searching (Kolling et al., 2012), the relative importance of healthiness relative to tastiness (Hare et al., 2011a) or the amount of reward necessary to overcome aversive experiences (Amemori and Graybiel, 2012). It is possible that pgACC

is involved in setting the threshold for such cost-benefit decisions: What is the amount of cost that I am willing to accept for a certain amount of reward? In light of this, one might assume that pgACC evaluates choices with a relatively stable and longer-term frame of reference.

There is, however, another body of work, mostly comprising human imaging studies, which implicate pgACC in social cognition. Previously I have described how in reward-guided decision making experiments it is difficult to distinguish pgACC activity from activity in adjacent vmPFC. Studies of social cognition often have the opposite problem: distinguishing pgACC from more dorsomedial frontal areas such as area 9. In an early imaging study, subjects judged whether adjectives presented on screen described: a) themselves; b) another person (George W Bush in this study); or c) were written in uppercase or lowercase (Kelley et al., 2002). PgACC was selectively active for self-judgments compared to the two other types of judgment. Note that, vaguely similar to the previous experiment, the experiment is not phrased as a decision task, but instead, again, a single trait is considered and a degree of acceptance/rejection is provided by the participant. These findings were subsequently corroborated by other studies, with the extension that judging trait descriptions for oneself and similar others increased pgACC BOLD signal while judging dissimilar others activated area 9 (Denny et al., 2012; Jenkins et al., 2008; Mitchell et al., 2006).

The co-occurrence of self-reflection signals and signatures of cost-benefit computations in pgACC is difficult to account for by reference to a single underlying mechanism. One suggestion is that, together with the hippocampus, pgACC plays an organizational role in memory retrieval and contextualization, particularly with respect to autobiographic memory (Milivojevic et al., 2015; Preston and Eichenbaum, 2013; Spreng et al., 2009). Quite why this would predict activation in social cognitive and value-guided decision making is not clear.

1.4.4. *Ventromedial prefrontal cortex*

Ventromedial prefrontal cortex (vmPFC) is the cortical region where probably most subjective value signals regarding objects and choice options have been found in humans, mostly using fMRI. The BOLD signal in vmPFC scales with the subjective value of food items (Hare et al., 2008; Plassmann et al., 2007). In a choice situation, vmPFC signals in some situations encode the sum of the presented options (Hare et al., 2011b), but in some other situations code the difference between the choice options with separable positive and negative effects of the chosen and unchosen value (Boorman et al., 2009; Philiastides et al., 2010). One way to explain the co-occurrence of these signals is that both types of signals are reflections of different stages of a decision-making process in which pools of neuron compete for choice (Hunt et al., 2012). VmPFC has a central role in decision-making even when the values guiding people's decisions are not ones that correspond to their own subjective preferences (Nicolle et al., 2012). In this sense, vmPFC activity is tightly linked to the choices that people actively pursue, rather than for instance only consider as a possibility. Such a view is compatible and possibly even predicted by observations that value-related vmPFC activity is increased for options that are attended compared to not attended, possibly even on a single neuron level (Lim et al., 2011; McGinty et al., 2016).

Electrophysiological recordings in monkey vmPFC are rare. As discussed above, the most likely monkey region corresponding to human vmPFC is area 14m. Studies from the laboratory of Padoa-Schioppa usually report recordings from a more lateral area of orbitofrontal cortex, which they tentatively label as area 13m (Cai and Padoa-Schioppa, 2014; Padoa-Schioppa, 2009, 2013; Padoa-Schioppa and Assad, 2006; Padoa-Schioppa and Assad, 2008; Xie and Padoa-Schioppa, 2016). Interestingly, one reason they initially focused on area 13m was that neurons there were sensitive to offer type, i.e. the identity of

a particular juice that a monkey could choose in the experiment. Strait and colleagues record from a more medial area in monkey orbitofrontal cortex that seems more likely to be area 14 and, therefore, to correspond to human vmPFC (Strait et al., 2014). They used a sequential binary choice task where options were characterized by reward probability and reward magnitude. They identified neurons that were sensitive to the expected value of the options and they also found that the value of the second option was coded relative to the value of the first option. This relative choice value or value difference coding is compatible with human fMRI studies showing value difference signals in vmPFC and further suggests that vmPFC is involved in deciding between competing options. Note that the abstract nature of the task precluded the identification of neurons sensitive to the identity (in terms of sensory properties such as juice type) of particular choice options. It is therefore unclear whether differences in the Strait and Padoa-Schioppa studies were due to differences in task design or whether they may indicate a relatively stronger involvement of vmPFC in decision making and more lateral areas in subjective valuation of objects and choice options.

1.4.5. *Summary*

Areas dACC, pgACC, area 9 and vmPFC, together with other regions, form a network that controls behavior. These regions have in common that they all carry value-related information, but differ with respect to the types, relative strengths and the functional roles of their value representations. In this thesis, I investigated the contribution of these areas to foraging-related behavior, in particular with respect to value tracking over time (Chapters 2,3) and social representations of self and others (Chapters 4,5).

VmPFC carries subjective value signals and its activity patterns precisely track economic decision processes, particularly when two or more options are offered and multiple attributes have to be integrated. VmPFC integrates the values of the choices that

are currently attended and that guide choice. These features could potentially enable vmPFC to compute goals for the organism that are then pursued over a longer period of time. It may be the case that the setting of such goals occurs by calibrating wide-spread network dynamics, rather than by the activity of a single brain region (Karlsson et al., 2012). VmPFC (but potentially also other brain regions) may trigger such recalibrations.

DACC's role in guiding choices and behavioral adaptation is, in contrast to vmPFC, particularly prominent when choices are pursued over a longer period of time and constitute a sequence towards a longer-term goal. Reward feedback is often coded in reference to these goals and whether changes in behavior are required. The framework in which choices or rewards are coded is rather one that relates to abstract modes of behavior such as exploitation/exploration or default/forage or stay/leave than one of specific goods that are chosen in binary decision making paradigms, inspired by behavioral economics, that have proved illuminating in studies of vmPFC and OFC. DACC activity might be related to maximizing average reward rates over the long run and in determining whether the current behavior is likely to achieve this. It might also relate to shorter-term goals in some tasks.

Value information in area 9 (or adjacent medial frontal areas in animal studies) seems to be less directly related to choice. However, there are striking indications that this area has access to surprisingly precise and flexible value representations. At the same time, area 9 has a role in learning about other people and when interacting with others. It might be that area 9 computes contextual variables regarding the choice setting that are not central to guiding the immediate choice (vmPFC might do that) or adjusting/controlling behavior with respect to longer-term goals (dACC might do that). A feature of such contextual representations might be that they are particularly flexible, i.e. can change very quickly. The value of, for example, food items are often relatively stable over time. They can be retrieved from memory. Other values are learned over many trials and behavioral

change often happens only after several pieces of evidence for it have been accumulated or when there are clear switching rules. By contrast, other people can have profound impact on individual success depending on the relationship to them and such relationships can change very quickly, for example when switching from cooperation to competition. Such a suggestion is compatible with ideas that social interactions are reliant on model-based computations (Lee and Seo, 2016)

PgACC has a role in cost-benefit decision-making, represents subjective value but does not seem to be primarily involved in the comparison of two or more choice options simultaneously. Across subjects, this region may predict the value of choices particularly well. Its parallel involvement in self-referential judgments is striking. It is hard to find a single common denominator to all the situations in which is activity is prominent. One feature of self-referential tasks that activate pgACC may be that subjects have to judge how likely they are to show a specific behavior/possess a particular trait. Hence, it may be that pgACC represents the values of actions with a particularly long frame of reference. Rather than computing whether an action or a choice is valuable in a particular situation, it might reflect the general value of acting in a certain way.

CHAPTER 2: Tracking the success of actions over time

2.1. Introduction

Many decisions are based on past experience because we expect the values of choices to be stable or to change only gradually. To choose effectively we must either track these values or, ideally, anticipate the future value a choice will have. While much is known about how choice values are learned from past experience, it is relatively unknown how people predict in advance not only that a value will change, but also how it will change. This problem lies at the heart of many judgment tasks, such as stock market trading or behavioral adaptation to the continuously changing reward rates animals and humans experience in foraging patches (Charnov, 1976; Stephens and Krebs, 1986). A simple way, however, to predict future choice values is to estimate the change in reward rates.

Previous studies have focused on decisions between two clearly defined and concrete options usually associated with specific stimuli (Chau et al., 2014; Hunt et al., 2013). While this is clearly an important decision-making mode, behavior often also reflects alternation between continuous engagement in the same behavior (for example when an animal repeatedly forages in one patch or a person maintains engagement with the same task) and behavioral change (when the animal moves to an alternative patch or the person switches task) (Charnov, 1976; Hayden et al., 2011b; Neta et al., 2015; Stephens and Krebs, 1986). In such situations it is critical to know prospectively how profitable the next repetition of an action will be.

We argue that it is possible to estimate future values of choices by comparing their recent value with the value they had further back in time. Identifying whether the choice's value is increasing or decreasing assists in identifying its likely future value. For some time there has been evidence that evaluation of a choice reflects comparison against

previous choice values. For instance, rats' approach speed to a given reference amount of food pellets depends on their previous reward experience. Having experienced more rewards on previous trials, rats are slower to approach the reference amount of food pellets; contrarily, having experienced fewer rewards beforehand, they are quicker to approach the same amount of food. These phenomena are known as positive and negative successive contrast effects, respectively (Crespi, 1942; McNamara et al., 2013). Similarly deciding whether to further commit to, or to quit, engagement in an action, the critical comparison is not between two stimulus values but between elements of the past reward history.

This chapter introduces a novel experimental paradigm that was used to investigate tracking of an option's reward rate over time. Human behavior in that task was analyzed in detail using computational modeling (Wittmann et al., 2016a). The subsequent chapter presents fMRI data acquired during the same experiment.

2.2. Methods

2.2.1. *Subjects*

Twenty-two subjects participated. One left the experiment as a result of claustrophobia. Another was excluded from data analysis due to excessive motion (final sample: 20 subjects; eight female; aged 21–32). All provided informed consent. The study was approved by the Ethics Committee of Oxford University (MSD-IDREC-C1-2013-095). Subjects received £20 as a show-up fee and a fraction of £15 depending on task performance.

2.2.2. *Experimental design*

We designed a reward learning task in which human subjects chose to further commit to, or to leave, a foraging-like patch based on its estimated future value. The patches were characterized by reward rate trends that could be discerned by comparing past and recent reward rates. This means that if the decision-maker knows the reward rates at two different time points in the past, then this is sufficient information to judge if the reward rate has increased or decreased overall between these time points. In environments with monotonic reward rate changes such knowledge can easily be exploited to extrapolate the reward trend, i.e. to predict the future value of the patch. The patches were derived from exponential reward rate curves similar to those in optimal foraging theory (Stephens and Krebs, 1986) (Fig.1a,b). The patches consisted of sequences of time steps on which either reward or non-reward events occurred (Fig.1c,d). The reward events were spread out such that their reward rates conformed to the underlying reward rate curves (Fig.1c). As such, the trials can be thought of as resembling fixed-ratio reinforcement schedules but where the ratio of actions to rewards changed monotonically over time. Subjects proceeded from time step to time step by pressing a button. At a predetermined time step, subjects

were offered a leave-stay decision (LSD; Fig.1d,e). For LSDs, subjects had to consider the 15 further time steps they would encounter after LSD, and decide, for this time period, whether to stay and further explore the environment they were in or to leave and re-engage with a pre-learned default environment with a stable reward rate. In essence, the LSD

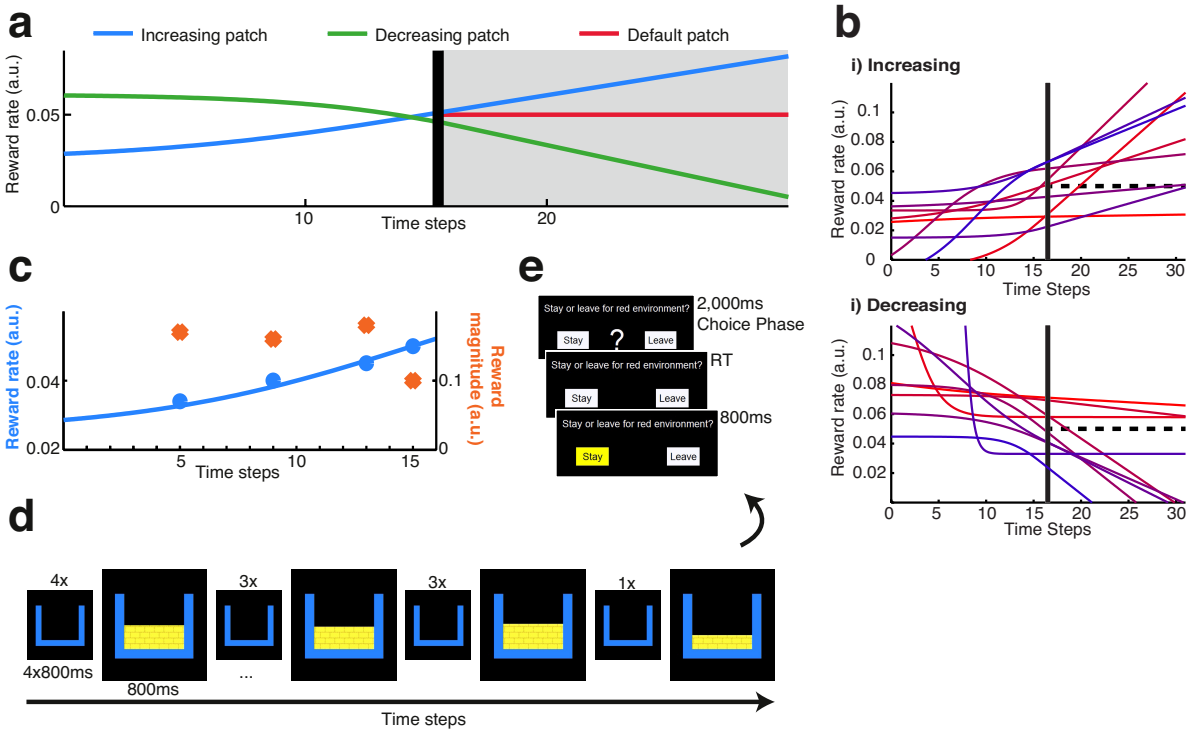


Figure 1. Experimental design and implementation. *(**a**) Two example patches with increasing (blue) and decreasing (green) reward rates. At leave-stay decisions (LSD; black) subjects chose between staying in the patch and switching to a default patch with a stable reward rate (red). In these examples, correct decisions (stay on blue, leave green) can be predicted from the reward rate curves. (**b**) All 18 reward rate curves from which trials were derived. Solid black line indicates LSD and vertical dashed line indicates reward rate of reference patch. For visualization purposes only, different coloring for reward rate curves was used and the curves were aligned so that the LSD is on the same time step. (**c**) Sequence of events corresponding to the blue reward rate curve before LSD in (a). Four reward events were presented at time steps 5, 9, 13 and 15. Their reward rates (blue dots), which conform to the reward rate curve (blue line), are calculated by dividing their reward magnitudes (orange dots) by the time delay from the previous reward event or the start of the patch. (**d**) Screen during events in (c). Empty boxes represent non-reward events, the height of "gold bars" in reward events represent their reward magnitudes. Each event was displayed for 800ms. Between events, a fixation cross was shown and subjects proceeded to the next event by pressing a button. Note that lastRR for this patch would be equivalent to the height of the gold bars in the final box divided by two time steps. (**e**) LSD followed, without time jitter, after the last reward event (boxes in the default environment were red, so it was labeled "red environment" for subjects).*

should be based on a comparison of the anticipated value of the current environment (the sum of rewards that would be delivered after LSD) and the pre-learned value of the default environment.

In each of 90 trials subjects proceeded through a patch consisting of reward and non-reward events (800ms each). Between events, a fixation cross was presented and button presses led to the next event. After time step 15, 16 or 17, subjects were offered the choice to stay for longer in the patch or to leave to a default environment with a known, stable reward rate. This meant that participants encountered LSDs at approximately similar positions in time on each occasion that they explored a new environment, however, the presence of some variability in LSD timing precluded precise anticipation of the LSD time by participants. Moreover, the analyses did not focus on activity that was simply linked to the main effect of LSD occurrence but to activity that was parametrically related to the reward experience prior to the LSD and which allowed subjects to make inferences about the reward likely to be received after the LSD. The design of the trials was based on 18 monotonic reward rate curves, nine increasing and nine decreasing, from each of which five unique sequences of reward/non-reward events were derived (Fig.1). The key manipulation was to assemble a set of patches such that the behavioral and neural effects of recent and past reward rates could be dissociated. The reward rate of a reward event was its reward magnitude divided by the number of time steps from the previous reward event or from the start of the sequence (time delay) (Fig.1c,d). The reward magnitude was indicated by the height of a golden texture ("gold bars") within a box presented on the screen. The time delay between reward events ranged between two to six time steps. On every trial, the LSD was, without time jitter, preceded by a reward event at the last time step of the sequence to keep the recency of the last reward event with respect to the LSD constant. Note that this meant that the last time step was always rewarded and the second last time step was never rewarded. Each LSD began with a 2,000ms "choice phase"

indicated by a question mark on screen (referred to as "LSD" in time course plots). Subjects then responded and their choices were highlighted in yellow for 800ms (Fig.1e). The left-right locations of “stay” and “leave” buttons were randomized. After LSDs, subjects continued through a sequence of 15 additional time steps. For stay choices, the underlying previous reward curve continued linearly based on its slope at the time of the LSD. If subjects decided to leave to the default environment, rewards were delivered at a fixed reference reward rate (i.e. the ratio of reward magnitude and time delay was identical for all reward events of the sequence) that subjects knew well from prior practice sessions. Although the specific event sequence differed between runs of the default environment, the sum of the rewards encountered there (leave value) was constant. For LSDs, subjects had to decide if the future value of the patch, the sum of reward encountered when further committing to it (stay value) was higher or lower than the stable leave value. On 66% of increasing patches, the stay value was higher than the leave value, while on 66% of decreasing patches, the leave value was higher than the stay value. Subjects received “bonus points” for making the better choice in proportion to the absolute difference between the stay and leave values and this determined the performance dependent monetary payoff subjects received at the end of the fMRI session. Optimal choices were defined as the ones with the higher payoff. Subjects received feedback about their accumulated bonus points after 1/4, 2/4 and 3/4 of the experiment. Finally, 60% of the trials were truncated after the LSD to shorten the experiment and to minimize choice feedback.

The generation of actual reward sequences from the theoretical reward rate curves was an iterative process. Beginning at the last time point of the curve (because the last event before a leave-stay decision or the end of a patch was always a reward event), a reward magnitude and a reward delay was randomly chosen under the condition that it conformed to the reward rate indicated by the reward rate curve at that time point. The

chosen reward delay determined the time point of the reward event preceding this event. For this, again, reward magnitude and reward delay was randomly chosen under the condition that it conformed to the reward rate indicated by the reward rate curve at that time point. The threshold for "conforming to the reward rate curve" was initially set to +/- 5% of the reward rate given by the curve, but it was expanded in 0.00001% steps in case no solution could be found. Note that for our analyses, it did not matter how closely the reward rates of the reward events in the sequence satisfied the reward rates originally intended for the curves, because all parameters used in our analysis were derived from the actual reward sequences and not from the theoretical reward rate curves. In other words, the reward rate curves that we show and use to guide all analyses of behavior and neural activity are the ones that were established empirically at the end of this iterative process.

2.2.3. *Training sessions*

Subjects were trained on the day before fMRI scanning (60-70 minutes) and on the day of scanning (15-20 minutes), immediately preceding the experiment. The full training session (only performed on the preceding day) comprised three parts.

First, subjects were familiarized with the default patch (Fig.2a, 10 trials). In each trial, alternatingly, subjects experienced a stable default patch and length-matched experimental patch. After both patches, subjects were asked to indicate which environment had contained more rewards (i.e. had the overall higher reward rate).

In part two, subjects experienced three patches (Fig.2b). They were told that a patch's payoff changed monotonically and that this becomes apparent when paying attention to both the reward magnitudes and reward delays of reward events. Subjects pressed through each patch three times with varying instructions. The instructions were to pay attention to 1) the change in reward magnitudes (ignoring delays) 2) the change of reward delays (ignoring reward magnitudes) and 3) the change in both magnitudes and

delay. This part contained no decision.

In the third part, subjects were asked to do a training version of the experimental task including leave-stay decisions (Fig.2c, 18 trials). Notably, they had the chance to experience the experimental patch or the default patch. The latter was only presented after the LSD when participants opted for the default option. Subjects were given performance

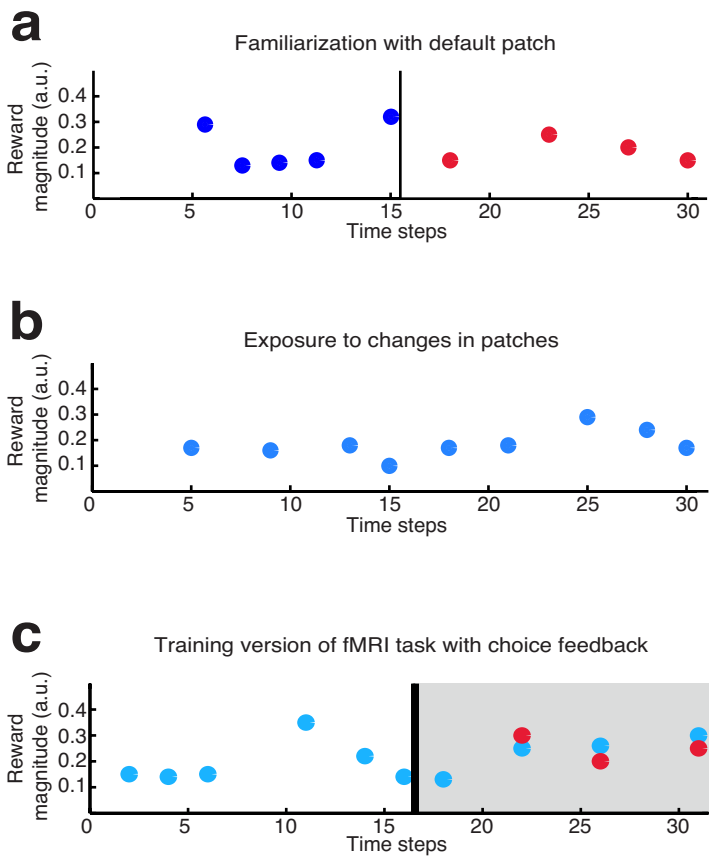


Figure 2. Three-step training session. *(**a**) Subjects were familiarized with the default patch. They experienced one experimental patch (blue) and the default patch (red) for 15 time steps and afterwards indicated which patch had earned them more rewards. Dots represent reward events of a given reward magnitude and the vertical line represents the cued transition to the other patch. (**b**) Subjects experienced some patches that changed in reward rate. The reward sequence displayed was taken from an increasing reward rate curve (note the increasing payoff of the reward magnitude to reward delay ratio: the reward magnitudes increase and/or the delays between rewards decrease) (**c**) Subjects performed a training version of the fMRI task with choice feedback. The vertical line represents the time of the LSD and dots in the grey area represent reward events from the experimental patch (blue) and the default patch (red). Subjects experienced only one or the other patch after LSD on any given trial. The reward sequence displayed was taken from a slightly decreasing reward curve (note the decreasing payoff of the reward magnitude to reward delay ratio: the reward magnitudes decrease and/or the delays between rewards increase).*

feedback after each trial. For performance feedback, subjects were told how many bonus points they had earned or missed depending on their choice. Bonus points were proportional to the overall payoff of the chosen compared to the unchosen patch after LSD. Hence, a positive number indicated a correct choice (earned points) and a negative number an incorrect choice (points missed). This step allowed subjects to discern which aspects of the reward environment were predictive of correct choices.

In sum, the training session gave subjects the opportunity to memorize the stable default patch, to grasp the structure of the reward environments and to learn how to make correct decisions. On the day of scanning, a shorter training session took place to refresh the experimental instructions. In the fMRI scanner, subjects had no opportunity to learn from direct choice feedback and 66% of trials were truncated after the leave-stay decision, discouraging changes in choice strategy during the fMRI session.

2.2.4. *Main reinforcement learning models*

To compare subjects' choices with a learning algorithm that integrates reward rates with a single time constant we devised a standard simple Rescorla-Wagner reinforcement learning (RL) model (RL-simple) (Sutton and Barto, 1998). More complex models were based on the formulas for RL-simple (see below).

At the beginning of a patch, a value estimate was set to the average reward rate experienced in the task up to that point and was zero for the first trial. The value estimate was then updated by the outcome of each time step of the patch sequence. Therefore, the number of value updates in a patch before LSD was equivalent to the number of time steps in the sequence. The value estimate was updated using a learning rate α fitted for every subject individually:

1) $Value(t+1) = value(t) + \alpha \times (outcome(t) - value(t))$

The size of the outcome was zero or positive and reflected the reward magnitude encountered at a time step. The value estimate at the time of the LSD (t = LSD), i.e. after the last event of the sequence, was used as the decision variable (DV):

$$2)\ DV = value(LSD)$$

The DV, representing the value of staying in a patch was compared with the value of the default patch to determine the model's choice. The probability of staying in a patch was calculated with a softmax equation:

$$3)\ P(stay) = \frac{\exp(\beta \times DV)}{\exp(\beta \times DV) + \exp(\beta \times value_{DEF})}$$

β is the inverse temperature of the softmax function and the constant $value_{DEF}$ represents the value of the default environment. Note that modeling $value_{DEF}$ as a free parameter means we do not use a RL mechanism to learn the value of the reference patch. Given that the reference patch is pre-learned and only very rarely encountered in the actual experiment (60% of trials end directly after the LSD), we determine the value each subject assigns to the reference patch empirically by treating $value_{DEF}$ as a free parameter that is stable over the course of the experiment. $Value_{DEF}$ was used in all RL models. We derived the choice probability from the stay probability on each trial:

$$4)\ P(choice) = \begin{cases} P(stay) & if\ stay \\ 1 - P(stay) & if\ leave \end{cases}$$

Overall, the free parameter set θ comprised α, β and $Value_{DEF}$. We fitted these parameters for every subject by minimizing the negative log likelihood (nLL) over all trials N, given a set of parameter values:

$$5)\ nLL = -\sum_{n=1}^{N} \log(P(choice(t)|\theta))$$

In further analyses, we expanded RL-simple to account for the reward rate trend guided choices observed. For all expanded models (except illustrative models, see below),

we fitted free parameters for each subject individually. The Softmax function used including the use of $value_{DEF}$ (equation 3) and fitting (equations 4 and 5, except the precise parameter set θ) were identical to RL-simple.

For the simplest extended model, RL-simple+lastPE, all features of the RL-simple+lastPE were the same as in RL-simple, except the DV. The DV included in addition to the last value estimate the prediction error (PE) of the last time step weighted by an additional free parameter (PEweight):

6) $DV_{simple+lastPE} = value(LSD) + PEweight \times PE(LSD-1)$

In sum, the free parameters for RL-simple+lastPE comprise α, *PEweight*, β and $value_{DEF}$.

Lastly, we constructed an average reward rate learner (RL-avgRR) that consists of two hierarchically organized RL mechanisms. First, we used a standard RL mechanism identical to the RL-simple model described above ($RL\text{-}avgRR_{part1}$). The value estimate (value) of $RL\text{-}avgRR_{part1}$ is updated at every time step using a learning rate α and a prediction error (PE):

7) $PE\ (t) = outcome(t) - value(t)$

8) $Value(t+1) = value(t) + \alpha \times PE(t)$

A second component of the new model, a parallel RL mechanism ($RL\text{-}avgRR_{part2}$) learns outcomes attenuated by value using an $PE_{expected}$ (contrasting effectively recent and longer-term past rewards), which can also be seen as learning prediction errors:

9) $PE*(t) = PE(t) - PE_{expected}(t) = outcome(t) - value(t) - PE_{expected}(t)$

10) $PE_{expected}(t+1) = PE_{expected}(t) + \alpha \times PE*(t)$

$PE_{expected}$ at the time of choice (i.e. after the last time step) was used in the DV:

11) $DV = PE_{expected}(LSD)$

Note that for the last PE* of $RL\text{-}avgRR_{part2}$ before choice, we fitted a separate learning rate as an additional free parameter (α-*lastPE*) allowing the model to calibrate the relative

weight of past relative to the most recent reward history. This aspect of the model was similar to the separate effect of the last PE in RL-simple+lastPE. This was done to account for particularities in our outcome sequences (a reward outcome was never received at the second to last time step while a reward outcome was always received at the last time step) and to make the RL-avgRR model comparable to RL-simple+lastPE model. Similar to the other models, at the beginning of each trial, both value and $PE_{expected}$ were set to the average reward rate experienced in the task up to that point. We used the same free parameter α for the learning rate of the basic and higher-order RL mechanism in the model. Therefore, the free parameters for RL-avgRR comprise α, *α-lastPE*, β and $value_{DEF}$.

We use an illustration to show the temporal integration function of RL-avgRR and compare it with RL-simple. We derived the functions that define the weights of past PEs and outcomes in $PE_{expected}$ at the time of choice, (i.e. the DV of RL-avgRR). For a standard RL model (for instance RL-simple), the value estimate can be mathematically decomposed into a sequence of n weighted past outcomes:

$$12)\ Value(t+1) = \sum_{i=0}^{n} \alpha \times (1-\alpha)^{i} \times outcome(t-i)$$

Considering that RL-avgRR learns based on simple prediction errors applying this formula results in ("PE" and "value" refer to components of RL-avgRRpart1 in the following equations 13-17):

$$13)\ PE_{\text{expected}}(t+1) = \sum_{i=0}^{n} \alpha \times (1-\alpha)^{i} \times PE(t-i)$$

For our illustration (Fig.6a), we show this function over a sequence of 15 time steps (n=15) using the median learning rate of avgRR (α=0.183, Table 1). For simplicity, we ignored the fact that the learning rate of $RL\text{-}avgRR_{part2}$ at the last time steps is decreased (median α-*lastPE*=0.134). To define the contribution of the outcomes rather than the PEs, the formula can be further broken down into:

$$14)\ PE_{\text{expected}}(t+1) = \sum_{i=0}^{n} \alpha \times (1-\alpha)^{i} \times (outcome(t-i) - value(t-i))$$

By using equation 12 with equation 14 this results in:

$$15)\ PE_{\text{expected}}(t+1) = \sum_{i=0}^{n} \left[\alpha \times (1-\alpha)^{i} - i \times \alpha^{2} \times (1-\alpha)^{i-1}) \right] \times outcome(t-i)$$

In Figure 6a-iii and 6a-iv we illustrate equation 15 and equation 13, respectively using the median learning rate of RL-avgRR. In figure 6a-i we illustrate equation 12 for the median RL-simple learning rate of 0.471. In figure 6a-ii the weights equal the learning rate of RL-simple (but normalized, see below).

As a supplement to the functions in Fig.6a describing the temporal integration dynamics of RL-avgRR and RL-simple, we ran two RL models, fitted on the whole group of subjects for illustration (grey bars in Fig.6a). Both models are identical to RL-simple and only deviate with respect to the calculation of the decision variable. The models explain choices reduced to the weighted history of PEs and outcomes, respectively. For the PE illustration model (Fig.6a-ii,iv), the DV was calculated purely on weighted PEs where the same free parameter weight was assigned to sets of three PEs (the last three PEs before LSD are assigned the same weight, etc.) using five free parameters $w_{1\text{-}5}$.

$$16)\ DV_{PEillustration} = \sum_{i=1}^{5} \sum_{j=1}^{3} w_i \times PE(LSD - (i-1) \times 3 - j)$$

Grey bars in Figure 6a-ii and 6a-iv show the fitted parameters $w_{1\text{-}5}$. The illustration of the outcome history weights (grey bars in Fig.6a-i and 6a-iii) was run in the same way using the DV:

$$17)\ DV_{outcomeillustration} = \sum_{i=1}^{5} \sum_{j=1}^{3} w_i \times outcome(LSD - (i-1) \times 3 - j)$$

Grey bars in Figure 6a-i and 6a-iii show the fitted parameters $w_{1\text{-}5}$. For the two illustrative models as well as the PE and the outcome functions, all parameter weights were normalized by their absolute sum, for display purposes.

2.2.5. *Supplementary reinforcement learning models*

In order to understand the dynamics of the modified RL model, RL-avgRR, more thoroughly, I ran a number of supplementary models. These models are not conceptually different from RL-avgRR, as the other main models are, but they help elucidate the functional dynamics of RL-avgRR in estimating choice value. Rationale and features of the modified models were the following :

RL-avgRR-no-α-lastPE: This model is identical to RL-avgRR, but does not use a separate learning rate for the last PE*. Hence, it has one less free parameter than RL-avgRR.

RL-avgRR-2Alphas: This model is identical to RL-avgRR, but separate learning rates were fitted to RL-avgRR$_{part1}$ (α-part1) and RL-avgRR$_{part2}$ (α-part2). Note that, as in RL-avgRR, the model contained an additional learning rate for the last PE* of RL-avgRR$_{part2}$ (i.e. PE*(LSD-1)) but not for the last PE of RL-avgRR$_{part1}$.

RL-avgRR-valuepart1: This model is identical to RL-avgRR, but the value estimate of RL-avgRRpart1 at the time of choice (value(LSD)) is also included in the decision variable of the model, weighted by an additional free parameter *weight-valuepart1* (Decision variable = $PE_{expected}$(LSD) + *weight-valuepart1*×value(LSD)). The rationale for the model is related to one feature of RL-avgRR that might seem counterintuitive — that the decision variable of RL-avgRR $PE_{expected}$, reflects an expected prediction error, and that it does not include an additional non-contextual estimate of the reward environment. Adding such a non-contextual estimate can be done in a very simple

way by including the simple value estimate from RL-avgRR$_{part1}$ at the time of choice in the decision variable, which is exactly what was done in this modified model.

RL-avgRR-freestart: This more exploratory model is identical to RL-avgRR, except that it uses an additional free parameter (start- $PE_{expected}$) as starting value of $PE_{expected}$ on every trial. As the prediction errors were calculated at the beginning of a trial in reference to the initial value one could imagine that this value affects the model fit. Thus, we made a variant of RL-avgRR, which finds the optimal starting value for each subject.

2.2.6. *Behavioral and reinforcement learning model analysis*

We used two complementary parameterizations of reward history. Our first behavioral GLM (Fig.4a) comprised the reward rates in discrete time bins moving back in time relative to the LSD. For each of five reward bins, the rewards received on three consecutive time steps were averaged. So overall, the reward bins covered 15 time steps, which was the minimum length of a patch. This resulted in five parametric regressors that represented reward rates for each of five time bins prior to LSD:

Behavioral GLM1: LSD-13-15, LSD-10-12, LSD-7-9, LSD-4-6, LSD-1-3

where LSD-1-3 represents the reward rate of the bin corresponding to the last three time steps prior to LSD and LSD-4-6 represents the reward rate of time steps 4-6 prior to LSD, etc. For this and all other behavioral GLMs, regressors were normalized (mean of zero, standard deviation of 1).

Our second set of parameterizations of reward history separates past rewards in lastRR and avgRR as well as lastRR-avgRR. LastRR is the reward rate of the last reward event, i.e. the magnitude of the last reward event divided by the number of time steps to the second last reward event. AvgRR is the sum of reward magnitudes of all reward events

in a patch (including also the last reward event) divided by the number of time steps until the LSD. Hence, avgRR includes lastRR. We used the following GLMs:

Behavioral GLM2: lastRR, avgRR

Behavioral GLM3: lastRR-avgRR

Behavioral GLM4: lastRR, lastRR-avgRR

The two GLM analysis approaches are, however, complementary and so there are relationships between them. An illustration of the relation between the time bin GLM (GLM1) and the lastRR/avgRR GLM (GLM2) is provided (Fig.3). LastRR shares more variance with more recent time bins compared to earlier ones while, by contrast, avgRR is increasingly correlated with earlier time points. However, the change is gradual, reflecting the continuous structure of the reward rate curves. A learner that based choices on the reward trend is predicted to weight lastRR positively and avgRR negatively. Such a choice pattern would translate into positive effects of recent time bins and a gradually decreasing effect of longer-term past time bins that eventually turns negative (compare with Results section).

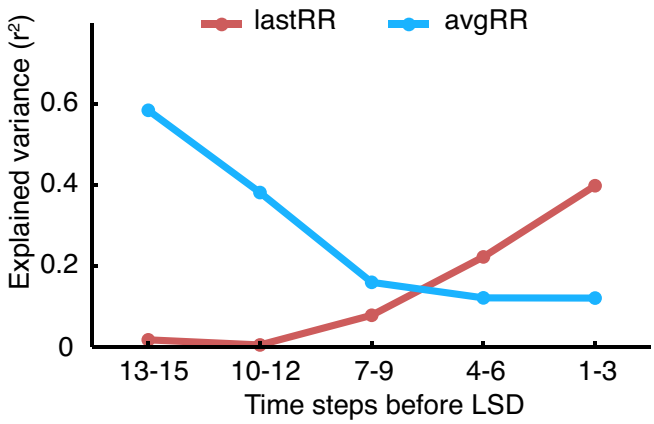


Figure 3. Illustration of the relation between the time bin GLM (GLM1) and the lastRR/avgRR GLM (GLM2). *Each data point represents shared variance (r^2) between a reward rate in the time bin-based GLM and lastRR and avgRR, respectively. Note that the underlying correlation coefficients are positive in all cases. Compare also with Figure 1b that shows that reward rate curves differ more early in a patch as a function of their trajectory compared to later. This asymmetry makes it easier to dissociate lastRR from avgRR in behavioral and neural analyses, because it reduces their correlation.*

All four behavioral GLMs were applied to subjects' actual choices and RL-predicted choices. Essentially, we compared the stay probabilities from the RL-model (p(Stay) from equation 3) with subjects' actual stay rates. We did this for RL-simple (Fig.4) and RL-avgRR (Fig.6b). To use the GLMs on RL-predicted choices, for every subject, we simulated binary choice behavior from the RL stay probabilities via a binomial distribution and then applied a GLM to the simulated choices. We repeated this binary choice simulation and application of GLM 1000 times per GLM and subject and averaged the resulting beta weights per subject to determine the influence of the regressors on RL-predicted choices. Error bars on the RL's beta weights always indicate standard error between subjects. For Figure 4d and Figure 6b-iv, we binned trials by the categorical type of reward rate trend (increasing or decreasing) and the optimal choice (stay or leave), calculated the actual stay frequencies and compared them with subjects' RL-predicted stay probabilities (p(Stay) from equation 3) averaged per bin. Note that in 66% of increasing trials it was optimal to stay while in 66% of decreasing trials it was optimal to leave (optimal in the sense of maximal payoff).

Behavioral data was analyzed with SPSS and MATLAB, using paired t-tests and repeated measures analyses of variance (ANOVA), including Greenhouse-Geisser correction where appropriate.

2.3. Results

2.3.1. *Opposing effects of recent and past rewards on choice*

A simple and descriptive way to demonstrate opposing effects of recent and past reward history on behavior is to measure the influence of rewards occurring at different times during a trial on the LSD using a general linear model (GLM). Therefore, we divided the reward history into reward rate bins reflecting reward received over five sets of three time steps each moving backwards in time from the LSD (LSD-1-3, LSD-4-6, etc.; Fig.4a). In our experiment reward rates in these different time bins share less than 25% of their variance so we can estimate their influences on choice behavior. Note also that this means that we could also test if choices were solely based on a patch's initial reward rates (correlation between initial and last time bin: r=0.02), which was not the case. In a repeated measures analysis of variance (ANOVA), we found a main effect of time bin ($F_{4,76}$=62.5; p=3*10^{-23}). The effect of reward on choice was increasing over time (linear effect: $F_{1,19}$=98.9; p=6*10^{-9}): While subjects tended to stay in a patch when reward rates in recent bins were high (LSD-1-3: t_{19}=10.12; p=4*10^{-9}; LSD-4-6: t_{19}=4.98; p=8*10^{-5}), the more distant the reward rate bins were, the more negative the effect of high reward rates on the decision to stay (LSD-10-12: t_{19}=-6.85; p=2*10^{-6}; LSD-13-15: t_{19}=-6.99; p=10^{-6}).

We compared subjects' choices with choices made by an individually fitted simple Rescorla-Wagner reinforcement learning (RL-simple) model(Sutton and Barto, 1998). We used the value estimate of the model at the time of the LSD (i.e. after observation of the last reward outcome) as decision variable. As such, RL-simple's value estimate reflects aggregate information about all rewards observed prior to choice and uses the same information available to subjects at LSD. We applied the former GLM to RL-simple simulated choices (Fig.4a, green bars). RL-simple captures recent positive effects of rewards (LSD-1-3: t_{19}=6.44; p=4*10^{-6}; LSD-4-6: t_{19}=7.58; p=4*10^{-7}), but is unable to

simulate a negative influence of past reward rate bins on choices to the same degree as seen in human subjects (paired t-tests on last two bins: LSD-10-12: t_{19}=8.11 p=1*10^{-7};

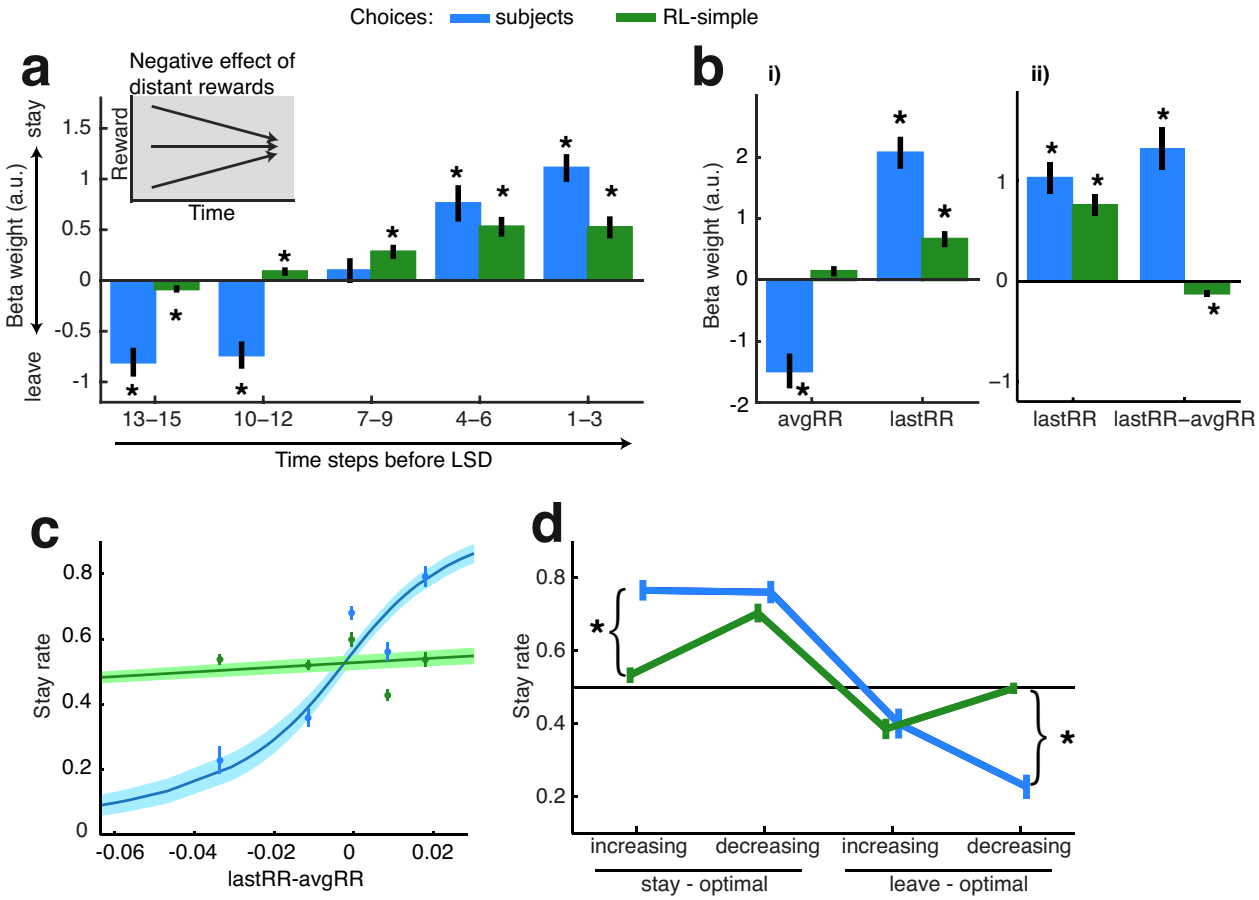


Figure 4. Recent and past reward rates influence choice in an opposing fashion. *(**a**) GLM predicting stay choices. Binning the reward history into bins of three time steps relative to the LSD (LSD-1-3, LSD-4-6, etc.) revealed the influence of discrete previous time steps on choice. Opposing effects of recent and past rewards emerged gradually in subjects' behavior (blue). A simple RL model (green) captured positive effects of recent rewards but failed to represent distant rewards as negatively as subjects did. Inset: The reason for the negative impact of early rewards becomes particularly clear when keeping recent rewards stable. The reward trajectory points more upwards the lower its starting position. (**b**) GLM predicting stay choices (same y-axis for i and ii). As in panel a reward rate trend guided behavior can be explained by a positive effect of a patch's most recent value (lastRR) in combination with a negative effect of reward rates in the past (avgRR) (i). Alternatively, this can be constructed as a positive effect of lastRR-avgRR and lastRR (ii). RL-predicted choices (green) captured part of the positive influence of lastRR but failed to represent longer-term rewards in the way subjects did. (**c**) Softmax functions of subjects' actual (blue) and RL-simple predicted (green) stay rates plotted against lastRR-avgRR illustrates that, overall, subjects' choices were influenced by the reward rate change in contrast to RL-simple. Overlaid are binned actual and RL-predicted stay rates. (**d**) Stay rates plotted by optimal choice and categorical reward rate trend. The simple RL model's choices (green) were close to random when the reward rate trend was predictive of the optimal choice (stay/increasing and leave/decreasing). It performed similar to subjects (blue) when the reward rate trend had to be ignored. (*, p<0.005 ; error bars are s. e. m. between subjects).*

LSD-13-15: t_{19}=6; p=9*10^{-4}). The strongly negative effects of rewards in past time bins in our human subjects, (LSD-10-12, LSD-13-15) resemble successive contrast effects in animals (Crespi, 1942) where the value of continuing to engage with a choice is lower when rewards had previously been higher (negative contrast effect), but continuous engagement is promoted when previous rewards were lower (positive contrast effect; inset Fig.4a).

Note that the temporal gradient in Figure 4a can be approximated using only two parameters: the reward rate estimated from the last reward event (lastRR) before the LSD, and the average reward rate (avgRR, including the last reward event) throughout the whole period before the LSD (Fig.4b,c). These two regressors capture the negative and positive effects of past and recent rewards on human behavior (lastRR: t_{19}=9.56; p=10^{-7}; avgRR: t_{19}=-6.23; p=5*10^{-6}). While the RL-simple model explained the lastRR effect (t_{19}=7.78; p=3*10^{-7}) it falsely predicted a positive instead of a negative effect for avgRR (t_{19}=3.53; p=0.0023; see Fig.3 for a description of the complementarity of Fig.4a and Fig.4b). Calculating the difference between lastRR and avgRR (lastRR-avgRR) results in a measure of the reward rate trend, accounting for both increasing and decreasing reward rate trends (positive and negative values of lastRR-avgRR respectively; Fig.4b,c). Instead of a positive effect of lastRR and a negative effect of avgRR, subjects' behavior can alternatively be captured by a positive effect of the lastRR-avgRR (t_{19}=6.23; p=5*10^{-6}), i.e. the reward rate trend, and lastRR (t_{19}=6.64; p=2*10^{-6}). RL-simple (green) predicted correctly that choices would be guided by lastRR (t_{19}=7.03; p=1.1*10^{-6}). However, it failed to reflect that subjects tended to stay when reward environments had become better over time (negative effect of lastRR-avgRR; t_{19}=-3.65; p=0.0017). As our investigations focused on the use of any type of monotonic reward trend in choice, lastRR-avgRR assumes the simplest type of trend, a linear one, rather than, for example an exponential trend. Although the underlying reward rate curves are exponentials, the sparseness of

reward events made it unlikely that subjects estimate very nuanced reward trend shapes. Note however, that subjects based their choices not only on lastRR-avgRR, but also on the absolute size of lastRR (Fig.4bii). In combination, the GLMs (Fig.4a,b) suggest that independent of a specific parameterization of reward history subjects weighted recent and past reward rates in an opposing fashion when making choices about future values.

The shortcomings of the simple RL model become particularly clear when binning the trials by their categorical reward rate trend (increasing or decreasing) and sorting them by optimal (defined by maximal payoff) choices (Fig.4d). When optimal choices had to be made against the reward rate trend (decreasing/stay-optimal and increasing/leave-optimal), RL choice predictions were close to the choices observed in subjects (although for decreasing/stay-optimal trials, we found a small but significant difference t_{19}=2.44; p=0.025). In these cases, correct choice could be based on lastRR alone. However, when optimal decisions depended on the reward rate trend (increasing/stay-optimal and decreasing/leave-optimal), human behavior was strikingly more optimal than predicted by RL-simple (t_{19}=9.72; p=8*10^{-9} and t_{19}=-10.31; p=3*10^{-9}, respectively). These results corroborate earlier results (Fig.4b) that subjects base their choices on both the instantaneous reward rate at the time of choice and the reward rate trend. Note that it was optimal to stay with increasing patches on 66% of occasions but to stay with decreasing patches on only 33% of occasions.

In sum, we show that subjects' choices were influenced by recent and past reward rates in an opposing fashion. Because a simple RL model integrates historical and recent rewards into a single estimate of an option's value, it does not represent recent and historical rewards with opposing weights and, therefore, does not make choices like the human subjects.

2.3.2. *Estimating prediction errors enables reward trend guided choice*

There are many possible modifications of simple RL algorithms. Prediction errors (PEs; differences between the reward outcomes that choices actually led to and prior reward expectations) are a simple measure of the option's reward rate change. Because PEs have a role in value updating they have been linked to learning (Sutton and Barto, 1998) but it is also possible that PEs may be used as a decision variable to guide decisions (Dayan, 2009). We tested whether an RL model that computes an expectation of PEs is able to model reward trend guided choice.

We modified a simple RL model based on a type of average reward rate model that has recently been described (Dayan, 2009). The computation of the average reward rate offers an estimate of the reward opportunities of an environment (Niv et al., 2007). Often the average reward rate reflects the opportunities that are currently foregone by engaging in one particular choice, a proxy for the value of leaving a patch (Constantino and Daw, 2015). However, in many natural environments choice values change over time. For example, when foraging in one particular tree, it may be possible to select only one type of fruit but the value of that foraging choice will change over time if the rate at which the fruit is encountered changes over time. In such a setting, the average reward rate does not reflect the value of many concurrently alternative options, but instead reflects the longer-term average value of the currently exploited option. Comparison of the option's recent value against its longer-term average value (McNamara et al., 2013) produces a simple measure of reward rate change. From an RL perspective, this is captured by calculating the difference between an option's instantaneous outcome and its expected outcome - the prediction error.

The modified model, RL-avgRR, consisted of two hierarchically organized RL learning mechanisms (Fig.5a). The first one (RL-avgRR$_{part1}$; Fig.5a top) was a standard RL

model (as RL-simple) and as such computed (standard) PEs on every trial. The second learning mechanism, RL-avgRR$_{part2}$ (Fig.5a bottom) used an estimate of PEs, PE$_{expected}$, which was updated on every time step using a higher-order prediction error, PE* (i.e. the difference of observed PE and expected PE). Over time PE$_{expected}$ learned a recency-weighted estimate of PEs. PE$_{expected}$ at the time of the LSD was used as decision variable.

We compared the goodness of fit of the RL-avgRR model with RL-simple. Moreover, we tested whether there is evidence that subjects' indeed used a longer-term estimate of prediction errors rather than only a single, most recent PE. For this, we

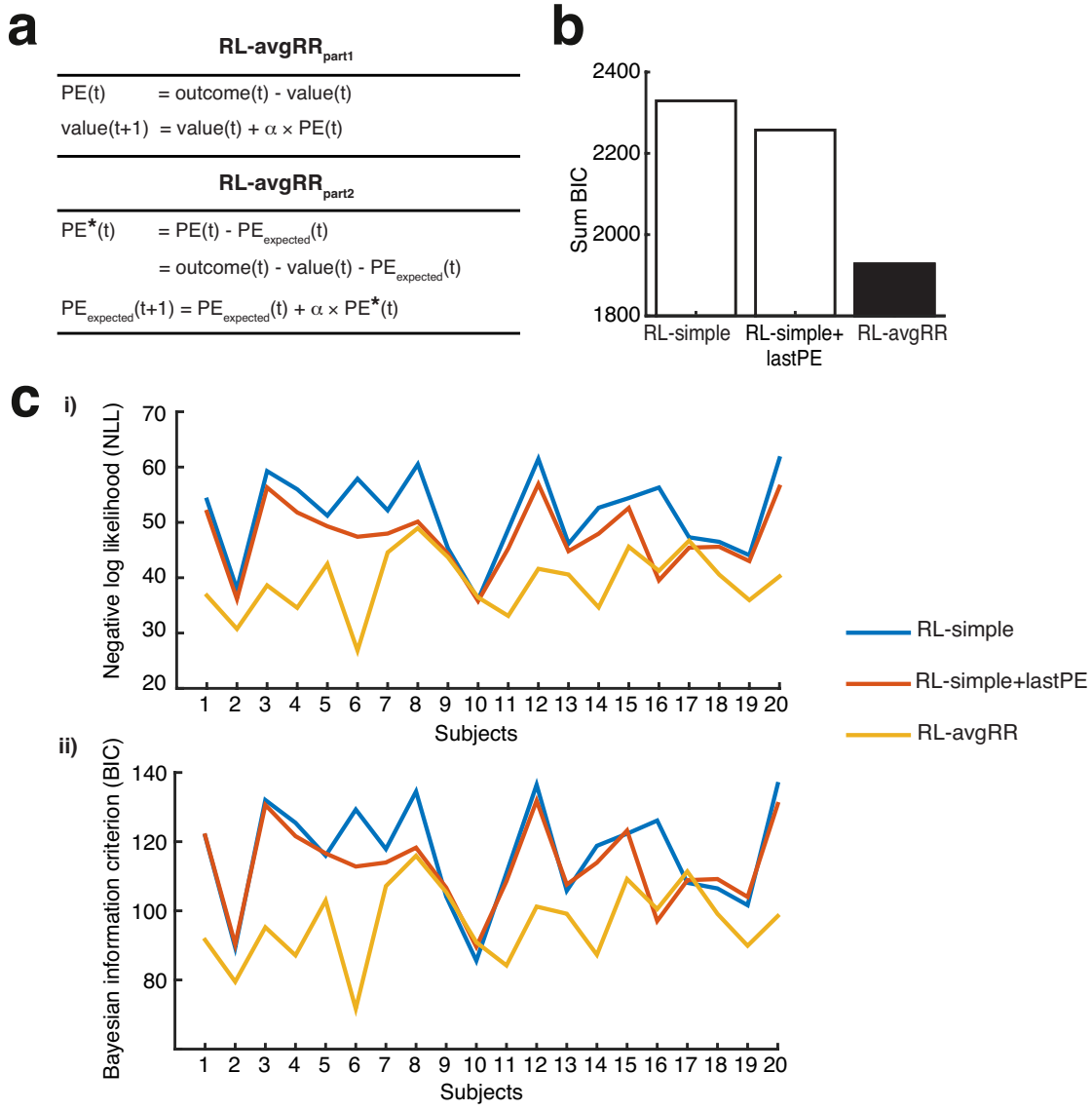


Figure 5. RL-avgRR equations and model comparison. *(**a**) Equations used in RL-avgRR. (**b**) Summed BIC scores for RL-avgRR were lower than RL-simple and RL-simple+lastPE. (**c**) Negative log likelihoods (i) and Bayesian information criterion (BIC) scores (ii) of fitted models shown for all subjects. Smaller values indicate better model fit.*

considered a more sophisticated version of RL-simple that uses, in addition to RL-simple's last value estimate, RL-simple's last prediction error (PE) to make a choice (RL-simple+lastPE; Methods). To investigate the goodness of fit, we calculated the Bayesian information criterion (BIC), which penalizes additional free parameters. RL-avgRR was by far the best of the three models (Fig.5b,c; lower BIC values indicate better model fit).

Computing expected PEs implies a very specific weighting of past outcomes and PEs. This weighting, i.e. the way events at different times in the past influence expected PEs, is consistent with the weighting of past events expressed in subjects' choices. It might be best explained in comparison with simple value estimates. For instance, given by the delta rule (Sutton and Barto, 1998), RL-simple's value estimate can be calculated as the recency-weighted sum of past reward outcomes (Fig.6a-i, green line). The weighting means that the value estimate varies more as a function of recent rewards than as a function of longer term past rewards, but all rewards have a positive weight of influence. This weighting profile is inconsistent with our finding that longer-term past outcomes have a negative influence on subjects' tendency to stay in a patch (grey bars in Fig. 6a-i on the left, also blue bars in Fig.4a). $PE_{expected}$, however, weights recent rewards positively and past rewards negatively (Fig. 6a-iii, orange line) similar to the weights expressed in subjects' choices; just as prediction errors are highest when encountering high rewards after experiencing low rewards beforehand (Bayer and Glimcher, 2005). The function expressed by the orange line is given by the learning rate and the model's equations alone. It is independent of other free parameters used in the model.

Expected PEs and simple value estimates can be decomposed into the weights of influence of past outcomes (Fig. 6a-i,iii). They can, alternatively, also be decomposed into the weights of past PEs (Fig. 6a-ii,iv). Also from this perspective, the weighting function of $PE_{expected}$ is qualitatively similar to the weights expressed in subjects' choices. The computation of expected PEs for choice, therefore offers a simple explanation of the

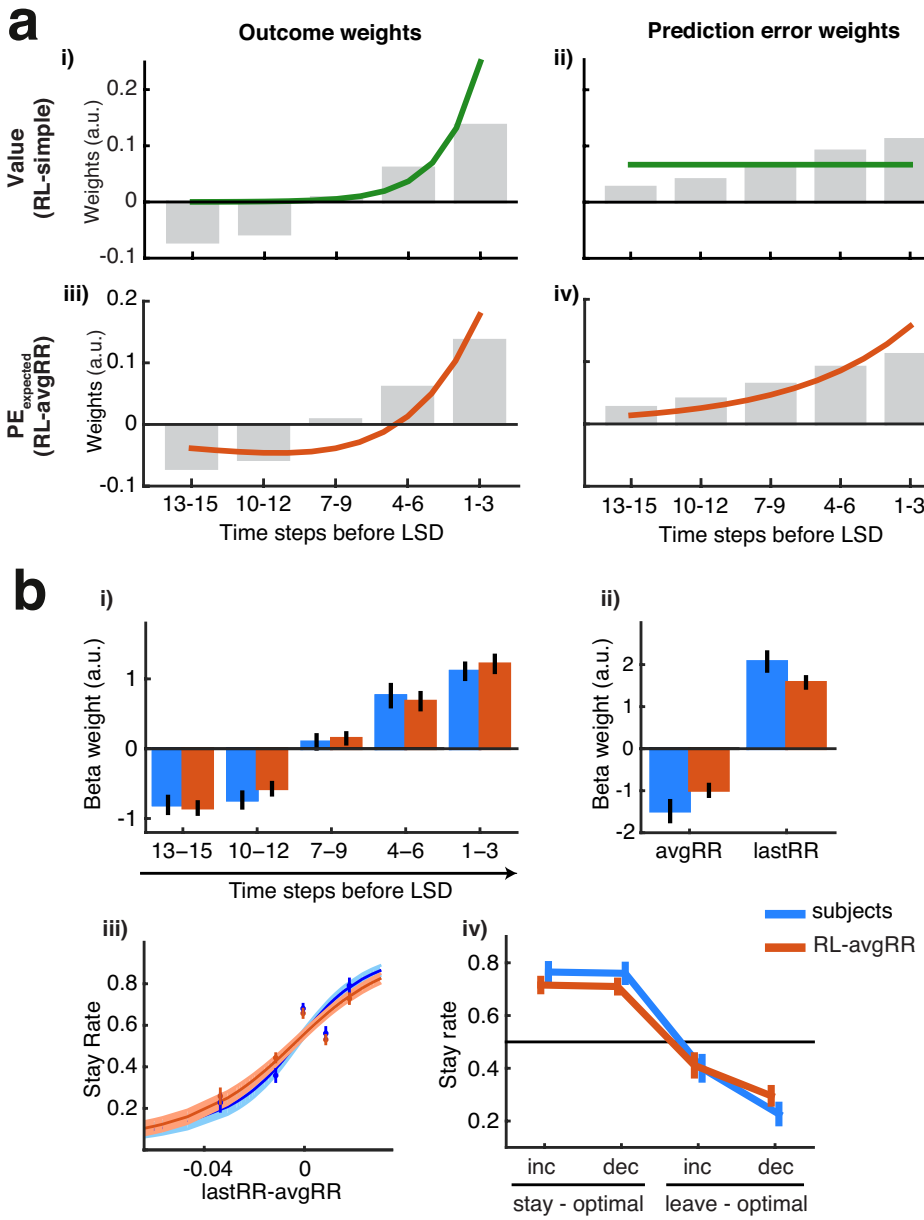


Figure 6. RL-avgRR model dynamics and choice predictions. *(**a**) Comparison of past outcome and prediction error (PE) weights used by a standard value estimate and* $PE_{expected}$ *at the time of choice. Grey bars indicate the weights of influence that past outcomes (i, iii, same data) and past prediction errors (ii, iv, same data) had on subjects' decision to stay in a patch. The lines indicate the amount of influence past events had on the calculation of a simple value estimate (from RL-simple; green line) and* $PE_{expected}$ *(from RL-avgRR; orange line). Note that the empirically determined weights correspond qualitatively to the weights used by RL-avgRR, but not to RL-simple. While simple value estimates are a recency-weighted sum of past outcomes (i), expected prediction errors are highest when encountering high rewards after initially poor outcomes (negative then positive weighting, iii). The same information as in i and iii can be presented as a function of PEs for RL-simple (ii) and RL-avgRR (iv). Again, the influence of past PEs on subjects' choices are qualitatively similar to the way* $PE_{expected}$ *is calculated from past prediction errors. (**b**) Analyses from Figure 4 were repeated for RL-avgRR. Unlike RL-simple (Fig.4), RL-avgRR made choice predictions (orange) that were similar to subjects' actual choices (blue). Note, in particular, how the weights in 3di mimic the theoretical weights shown in ciii and that RL-avgRR is able to represent past rewards negatively. See Figure 4 for legends.*

positive and negative time-linked reward effects that characterize reward trend-guided choice.

Applying RL-avgRR to our experimental schedule simulates the type of behavior we observed in subjects. We repeated four analyses we applied to RL-simple in Figure 4 using RL-avgRR instead. Just as seen in actual human behavior, RL-avgRR predicted a smooth transition in the negative to positive effects of more distant versus more recent rewards on choices (Fig. 6b-i). Analogously, this can be summarized as simultaneous positive and negative effects of lastRR and avgRR (Fig. 6b-ii). Consequently, RL-avgRR predicted choices in accordance with the reward rate trend (Fig. 6b-iii) and also choices in cases where it was optimal to follow the reward trend as well as the recent reward rate (Fig.6b-iv). Overall, these analyses show that, in contrast to RL-simple (Fig.4), choice values produced by RL-avgRR are negatively influenced by longer-term past rewards. This enabled the model to extrapolate reward trends.

In sum, a RL-based learning mechanism that tracks not only the longer-term history of encountered rewards, but also the longer-term history of predictions errors can be used to model reward-trend guided choice.

2.3.3. *Supplementary RL analyses*

In supplementary RL analyses, we aimed to describe the dynamics of RL-avgRR in more detail. These analyses comprised firstly additional model comparisons of related models that were modified versions of RL-avgRR. However, RL-avgRR proved to be the best fitting simple model. In a subsequent second step, we explain conceptually why RL-avgRR is sufficient to capture behavior in our experiment even in the absence of a simple value estimate in the decision variable.

In the modified RL-avgRR models, we changed a single aspect of RL-avgRR to evaluate how this change affected the model dynamics. We found that all RL-avgRR

models explained behavior in a robustly better manner than the two reference models (Fig.7a). This demonstrates that neither of the modified aspects in isolation is crucial for RL-avgRR to work in our experiment and suggests that the RL-avgRR's goodness of fit is indeed related to the model's core feature of calculating expected prediction errors (see also Table 1 at the end of this section for median parameter weights of main and supplementary models). To corroborate this finding, we calculated correlations of the model's decision variable averaged over subjects using the unsigned correlation coefficients (Fig.7b). As expected from the model comparison, the choice computations performed by all versions of the RL-avgRR model are highly similar (all correlations above 0.929) but distinct from the two reference models. In detail, this sheds light on the following features of RL-avgRR (discussed based on the modified models):

RL-avgRR-no-α-lastPE: The use of a separate learning rate for the last PE* was not crucial with respect to the model fit. The model had a comparable low BIC as RL-avgRR. One reason why the separate learning rate of RL-avgRR improves model fit slightly might be that the last time step before decision (the time step to which the additional learning rate relates) is always rewarded in our experiment. This was done to control for reward recency at the time of choice. However, for all other time steps in a trial reward events occurred much more sparsely.

RL-avgRR-2Alphas: We found that using different learning rates for RL-avgRR$_{part1}$ and RL-avgRR$_{part2}$ did not improve the model fit. Note also that the parameter weights for the two alphas were very similar (median weights of 0.173 and 0.185) explaining why the use of separate learning rates did not further improve model fit.

RL-avgRR-valuepart1: The rationale for the model is related to one feature of RL-avgRR that might seem counterintuitive — that the decision variable of RL-avgRR PE$_{expected}$, reflects an expected prediction error, and that it does not include an additional non-contextual estimate of the reward environment. Here, we added RL-avgRR$_{part1}$'s value

estimate to the decision variable. However, in our experiment this addition to the model did not improve the model fit further in terms of BIC, although a slight decrease of negative log likelihood in the model fitting was observed. This demonstrates that in the context of our study RL-avgRR's $PE_{expected}$ contains sufficient information to reflect both instantaneous reward and reward rate trend. However, to enable the model to be sensitive to absolute reward levels under more stable circumstances, it could include the additional simple value estimate.

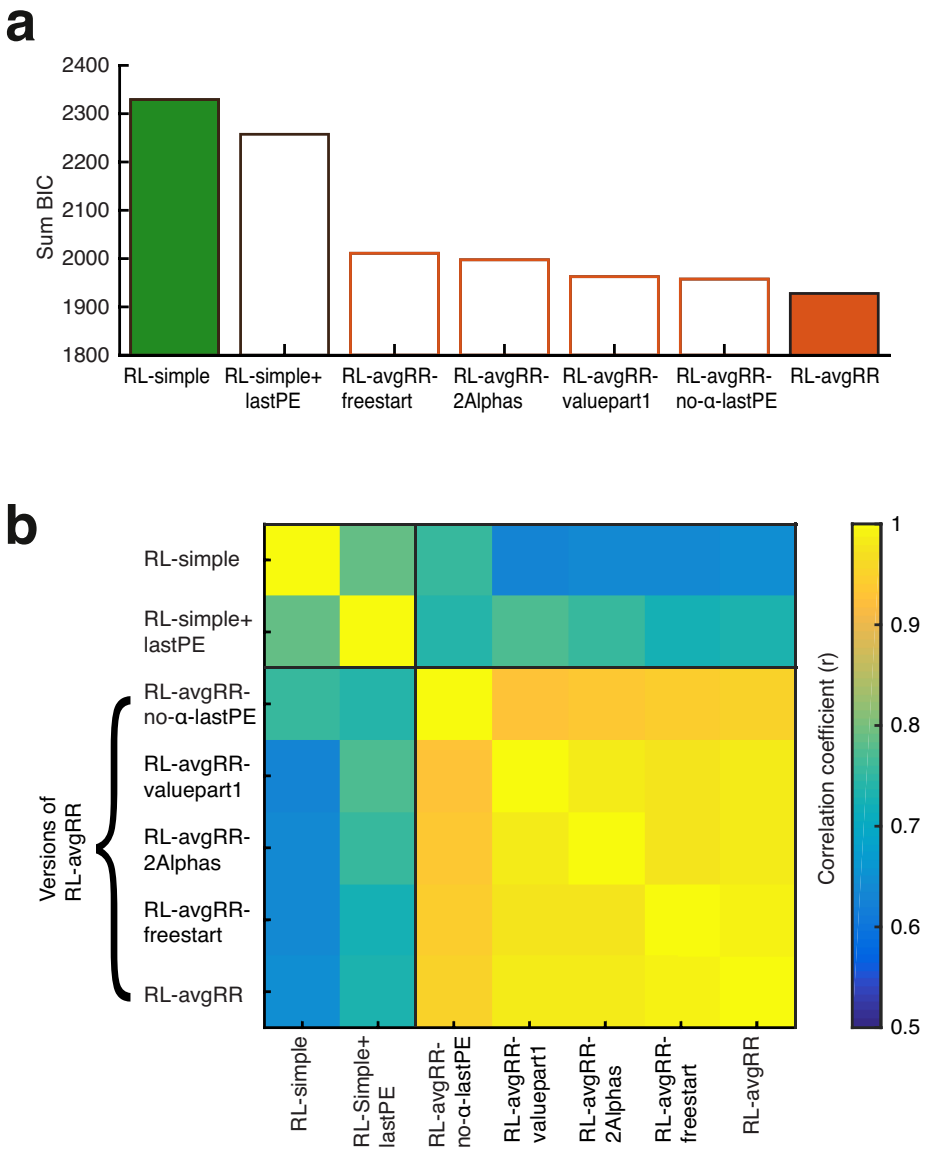


Figure 7. Extended RL model comparison. (a) *Supplementary model comparison including the three main models (RL-simple, RL-simple+lastPE and RL-avgRR) and four modified versions of RL-avgRR (white bars with orange borders). The group sums of the Bayesian information criterion (BIC) are shown for each model sorted by BIC value. Smaller values indicate better model fit. (**b**) Correlations of model's decision variables. Black solid lines differentiate versions of RL-avgRR from the reference models.*

RL-avgRR-freestart: As the prediction errors were calculated at the beginning of a trial in reference to the initial value one could imagine that this value affects the model fit. Thus, we made a variant of RL-avgRR, which finds the optimal starting value for each subject. However, this did not improve the fit. In fact, it would be surprising if it had improved the fit, as the important aspect of the model was to differentiate between different environments, for instance ones that had initially high reward rates or initially low reward rates. This is further corroborated by the high correlation of the decision variable that is calculated with this model and with the original RL-avgRR (mean correlation of decision variables is 0.9916, Fig.7b).

Finally, we discuss a particular feature of RL-avgRR that may seem unusual: the fact that a simple value estimate is not included in its decision variable. The decision variable instead consists only of $PE_{expected}$ at the time of choice. While we show that a memory of previous PEs can be used to decide according to reward trends of the environment, PEs are also influenced by the absolute reward levels in a trial. Hence, it is not necessary to use a simple value estimate in the decision variable in addition. This can be explained in terms of the learning process in a trial, where information about absolute reward levels is initially reflected by the reward trend estimate (Fig.8a) or, in a complementary way, in reference to the precise pattern of the outcome weights used by RL-avgRR (Fig.8b). A critical feature for both explanations is the time horizon over which rewards are experienced. Due to the relatively small learning rates (median of 0.183) there is no need to add a simple value estimate to the model's decision variable (Fig.8c). This might well be different in other settings. In this sense, our findings indicate that PEs can play an important role in guiding choices directly (rather than only as a consequence of learning). This is corroborated by the finding that it can substitute effects of simple value estimates in our study. However, that simple value estimates often are the crucial decision variable guiding choice is obvious and of course not challenged by our findings.

Figure 8a shows the development of $PE_{expected}$ over time separated by both reward trend (increasing/decreasing) and optimal choice at decision point (stay/leave; the optimal choice in a trial is defined as the one that leads objectively to more rewards on the current trial). Note that rewards are presented intermittently in our experiment. This explains sudden increases or decreases in the temporal evolution of $PE_{expected}$, particularly during the last two time steps (a reward is always presented on the last time step and is never presented on the second to last time step, see Methods). We did this to ensure that recent information about reward rates was available at the decision point in a similar manner in both increasing and decreasing environments in which it was optimal to stay in which it was optimal to leave. If this had not been then done then whether reward was delivered at the last time step or not could have contaminated our model estimate. Hence, the important feature of this plot is the general development of the decision variable (i.e. $PE_{expected}$), rather than interpreting each time step separately. Early in a trial, $PE_{expected}$ is high or low as a function of the type of reward environment (increasing or decreasing), but later on as a function of the optimal choice (stay or leave; right hand side). In detail, early in a trial, $PE_{expected}$ reflects learning about the average reward levels of the environments, which are initially higher in decreasing patches. In other words, $PE_{expected}$ is high at first in decreasing patches as these are initially experienced as increasing as the agent first gathers experience of the environments. Vice versa, the increasing patches are experienced as decreasing relative to subjects' initial expectations. Expressed in terms of simple prediction errors, this means that early in a trial positive simple PEs predominate in the decreasing environments; the model learns that the current rewards are better than what is expected on average. This is how the model's $PE_{expected}$ initially captures the absolute reward levels of the environment. From that point on, additional rewards modify $PE_{expected}$ in the direction of the reward rate trend. Therefore, $PE_{expected}$ decreases more in patches that are more strongly decreasing (and leave is the optimal choice) compared to patches that decrease but

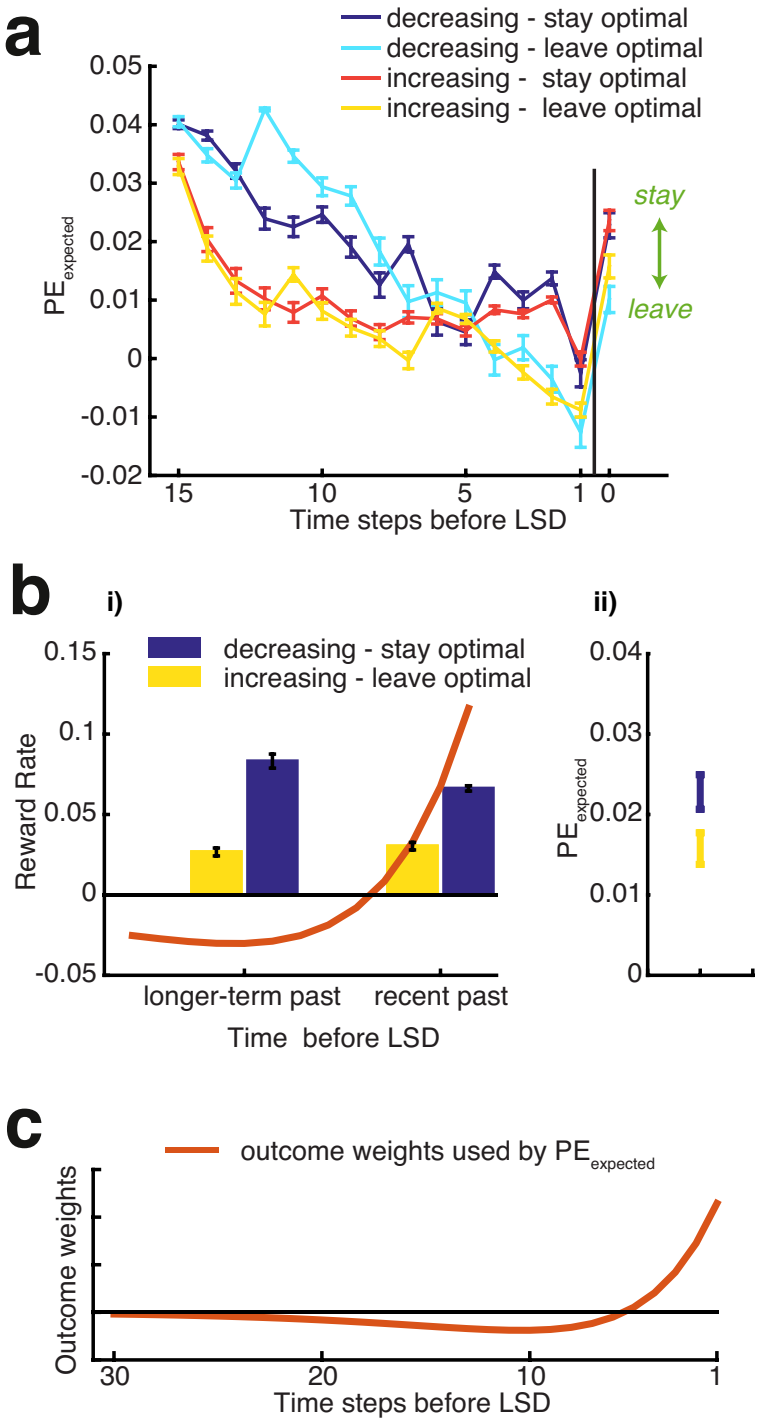


Figure 8. RL-avgRR and choices not based on the reward rate trend. *(**a**) The average evolution of $PE_{expected}$ over the course of a trial (left to right). $PE_{expected}$ at the time of choice is the decision variable of RL-avgRR and it is shown for four different trial types based on their reward trajectories (increasing or decreasing) and whether stay or leave is the optimal choice. (**b**) Reward rates early and late in a trial for trials where choices are to be made against the reward rate trend. The weighting function is shown in orange and illustrates that the positive weight of recent rewards is higher than the negative weight of longer-term past rewards (i). This translates in a higher decision variable ($PE_{expected}$) for decreasing/stay compared to increasing/leave trials even though decisions are to be made against the reward rate trend (ii). (**c**) The full weighting function of RL-avgRR over a longer time horizon. See main text for more explanation on the figure panels.*

not to such a strong degree (stay still optimal). In contrast, $PE_{expected}$ increases more strongly for patches with a stronger increase in reward rates (and stay is optimal) compared to patches were reward rates increase, but not sufficiently enough (and leave is optimal). In summary, RL-avgRR scales initially with absolute reward levels in a patch in terms of their positive or negative deviation from a longer time average reward rate (i.e. their initial trend). This estimate is then modified by the later increase or decrease of the reward rates in the patch, i.e. reward trends at decision point (increasing/decreasing). Therefore, the model's $PE_{expected}$ is related to both the reward rate trend as well as the absolute reward levels in our experiment.

Moreover, there are a few things to note in Figure 8a. First, subjects are naive with respect to the environment when they enter it. If this had not been the case and subjects did not have to learn initially about the reward levels in increasing and decreasing patches, then the early PEs could not be used to approximate absolute reward levels. However, as long as the starting value of the model reflects a relatively constant feature of the past reward history (i.e. is not initialized as the first reward rate of the specific patch), relative preferences based on developing trends and higher and lower absolute reward levels can be reconstructed in terms of the relative deviation from that initial constant. This is also why different starting values lead to similar model fit (see Figure 7). Second, note that as learning about the environments proceeds, all prediction error estimates ($PE_{expected}$) decrease. This effect is superimposed on the other effects described before (initial clustering of $PE_{expected}$ by increasing/decreasing, then by stay/leave). Third, the fact that rewards are presented intermittently in our experiment (as well as in most probabilistic learning tasks) leads to relatively low learning rates for RL-avgRR (median 0.183), reflecting integration over many time steps to get a stable value estimate.

Figure 8b shows a complementary explanation of why PEs also contain information about what we called absolute reward levels is related to the precise pattern of

outcome weights used by RL-avgRR in our experiment (orange line in panel i; from $PE_{expected}$ as in Fig.6a-iii). We illustrate this for trials of our experiment in which decisions have to be made against the reward rate trend, as those could not be made correctly without information about absolute reward levels. We show the average reward rates of reward events binned by recency (median split in recent past and longer-term past relative to choice) for trials in which it was optimal to leave although they were increasing (yellow) and in which it was optimal to stay although they were decreasing (dark blue) (Fig.8b-i). Optimality was defined as the choice that led to objectively more rewards on the current trial, as in Fig.8a. For these two trial types, decisions had to be made based on experienced reward levels and against the reward trends. That the model could do this is clearly reflected by the higher decision variable of RL-avgRR (i.e. $PE_{expected}$) for decreasing/stay compared to increasing/leave trials (Fig.8b-ii; see also Fig.4d). A complementary explanation to above is to illustrate the outcome weights directly. For this we need to look more closely at the model's positive (orange line, right hand side) and negative weights (orange line, left side) of recent and past outcomes, respectively. These weights are not balanced (i.e. they do not sum up to zero) but instead positive weights are stronger than negative weights within the time frame of approximately 16 time steps of each trial. Hence, there is a stronger positive than negative weighting (the relation is approximately 4:2.5 for positive compared to negative weights for the function shown in this figure). This distribution of weights is caused by an interaction between the time horizon studied and a relatively low learning rate. The stronger positive influence of recent compared to the negative influence of longer-term past rewards enables the model to differentiate between environments with varying absolute reward levels. In other words, even though there was a negative contrast effect that higher levels of past rewards exerted on the valuation of recent rewards, this contextualization was not complete (see panel c for a case where the

contextualization is complete and an additional simple value estimate is needed to guide choices).

The following numeric example gives an illustration how the imbalanced weight set can be used to differentiate between high and low reward environments in the absence of a reward trend. Imagine a stable environment where one reward was delivered in the negative past time window and one reward was delivered within the positive time window (for instance think of the yellow bars in the plot having both a reward rate of 1). Imagine also the positive and negative part of the orange line were just two values, a positive one and a negative one, summarizing the overall weight during that recent and longer-term past time period (for instance -2.5 and 4, as is actually the ratio for the orange line). $PE_{expected}$ could then be calculated as the (negative) sum of the past rewards times the negative past weight (-2.5) plus the recent rewards times the positive recent weight (4). In this case, $PE_{expected}$ of RL-avgRR would be 1*(-2.5) +1*4=1.5. In an environment with a higher stable reward rate (for instance the dark blue one), e.g. two rewards per time window, $PE_{expected}$ would be 2*(-2.5) +2*4=3. Consequently, $PE_{expected}$ and thus the decision variable is higher (3 compared to 1.5) for the environment with the higher stable reward rates (dark blue). Note that in this example the resulting estimates are positive, just as it was predominantly the case in our experiment (positive values in ii), reflecting the imbalance of stronger positive compared to negative outcome weights. This example illustrates how $PE_{expected}$ is able to differentiate between stable environments with varying absolute reward rates. In other words, the imbalance of positive and negative weights means that the RL-avgRR decision variable contains not only information about the later reward trend but also the non-changing part of the reward environment, as this components has not been completely balanced out. In other words, given the learning rate and the number of time points over which people experienced the environment, the outcome experienced early and late in the environment were not balanced in regard to the beta weights.

For completeness, we also want to outline example cases where it is likely to be important to include an additional simple value estimate in the decision variable to capture absolute reward levels (Fig.8c). These are likely to be ones where learning of the environment occurs over a very long time period (longer than in our experiment). To illustrate this, we show the outcome weights used by RL-avgRR's $PE_{expected}$ expanded over a longer time period (same function as shown in Fig.6a-iii, but longer time frame). Here, the negative tail of the function (orange) on the left hand side converges over 30 time steps. In this case, all outcome weights sum up to zero. In other words, positive recent and negative past weights cancel each other out. This means that $PE_{expected}$ in a completely stable environment would be zero, any decrease in reward rate recently would mean a negative $PE_{expected}$ and any recent increases would mean a positive $PE_{expected.}$ In such a setting, $PE_{expected}$ would only reflect the change in reward rates and not absolute reward levels. Note that even in this setting RL-avgRR would still be perfectly able to perform the computation that is the emphasis of this experiment; it would use a longer-term memory of PEs to estimate reward trends i.e. recent changes in the reward environment. Only a simple RL value estimate (which RL-avgRR computes anyways to estimate the PEs) would have to be included in the decision variable to capture differences in absolute reward levels. Therefore, in contexts with very stable environments longer than participants learning horizon or when the initial reward level is known to subjects preceding a trial, the inclusion of a simple RL's value estimate could allow a model to capture both trend and absolute reward level information. However, in our experiment $PE_{expected}$ alone, without a simple value estimate, was sufficient to dissociate between environments with rich and poor but relatively stable reward rates (see Fig.8a,b) as such reward levels had only recently been learned themselves.

Main models:

	RL-simple		
Parameter	α	β	$value_{DEF}$
Median (SE)	0.471 (0.059)	40.909 (6.239)	0.087 (0.008)

	RL-simple+lastPE			
Parameter	α	*PEweight*	β	$value_{DEF}$
Median (SE)	0.493 (0.071)	-0.308 (0.072)	57.960 (48528.45)	0.047 (0.009)

	RL-avgRR			
Parameter	α	α-*lastPE*	β	$value_{DEF}$
Median (SE)	0.183 (0.016)	0.134 (0.012)	240.46 (20.89)	0.015 (0.002)

	RL-avgRR (fitted on all subjects)			
Parameter	α	α-*lastPE*	β	$value_{DEF}$
value	0.184	0.163	161,14	0.017

Supplementary models:

	RL-avgRR-no- α-lastPE		
Parameter	α	β	$value_{DEF}$
Median (SE)	0.173 (0.014)	174 .643 (19.023)	0.020 (0.001)

	RL-avgRR-2Alphas				
Parameter	α-*part1*	α-*part2*	α-*lastPE*	β	$value_{DEF}$
Median (SE)	0.173 (0.037)	0.185 (0.048)	0.131 (0.021)	266.867 (31.398)	0.016 (0.003)

	RL-avgRR-valuepart1				
Parameter	α	*weight-valuepart1*	α-*lastPE*	β	$value_{DEF}$
Median (SE)	0.192 (0.017)	-0.030 (0.122)	0.139 (0.030)	236.565 (35.711)	0.017 (0.010)

	RL-avgRR-freestart				
Parameter	α	α-*lastPE*	β	$value_{DEF}$	*start*- $PE_{expected}$
Median (SE)	0.184 (0.016)	0.137 (0.013)	238.384 (20.987)	0.015 (0.002)	0.0233 (0.225)

Table 1. Summary of parameter estimates of RL models. (SE is standard error). *Note in particular the smaller parameter values for $value_{DEF}$ for versions of RL-avgRR compared to versions of RL-simple. The reason for this is that while RL-simple computes standard value estimates reflecting a recency-weighted average of the reward environment, the estimates computed by RL-avgRR reflect a contrast measure of recent and past reward rates. As such, they are the sum of positively and negatively weighted past rewards, resulting in numeric values closer to zero.*

2.4. Discussion

Considerable emphasis has been placed on the way both human and animal behavior is guided by value expectations based on learned associations between specific stimuli and reward outcomes. In many natural foraging situations, however, part of what animals are doing may simply be deciding whether to continue a behavior or whether it will be more profitable to change behavior (Charnov, 1976; Hayden et al., 2011b). This is particularly important in environments, in which no immediate cues act as instructions to stop or switch behavior (Isoda and Hikosaka, 2007; Mars et al., 2009; Neubert et al., 2010). Such cases are not only likely for foraging animals but also for human task-switching; we continue to engage in a behavior while it appears profitable, but disengage as the rate of return diminishes, for example in equity markets (De Martino et al., 2013b). In such situations a comparison of the current shorter-term rate of reward with the longer-term average reward rate of the environment is highly informative. It might allow the decision-maker, for instance, to make an anticipatory switching choice away from a currently exploited option just before the option's value falls under the value of the alternatives.

Trends in reward rates can readily be estimated if an agent weights recent rewards and a longer-term average in an opposing fashion (Fig.4). We show that this opposing weighting is not captured by value expectations calculated with a standard RL model. However, if the RL mechanism is able to compute an expectation of prediction errors, then this expected prediction error can be used to perform reward trend guided choice. Contrasting current outcomes with a longer-term average, PEs reflect a simple measure of reward rate change. In the same way as past outcomes influence subjects' choices, expected prediction errors at the time of choice will be highest when first experiencing small

rewards and only more recent rewards are high, i.e. reward trends are increasing (Fig.6a-iii).

Both our modified RL model as well as our behavioral analyses suggest that predicting the reward rate trend can be conceptualized as related to successive contrast effects (Crespi, 1942) in which past rewards have a negative effect on likelihood of continuous engagement in a choice. This highlights that estimating PEs is one way to track reward environments on various time scales and that reward trend guided choice can be sufficient to successfully maneuver through certain types of reward environments. However, the relative importance of expected PEs (particularly in comparison with a standard RL value estimate) critically varies depending on how useful reward trends are in guiding choice. Reward trends might not be useful in guiding choice, for instance, when changes in the reward environment are abrupt and unpredictable rather than slow and monotonic, or when reducing the time window over which future rewards need to be estimated. Although there might be other models that capture reward trend guided choice, using the expected prediction error to solve this problem is appealing for several reasons. It is parsimonious, because it requires only simple modifications of a standard RL model. Moreover, it is consistent with recent suggestions that prediction errors might not solely have a role in value updating. PEs have recently been linked to subjective wellbeing (Rutledge et al., 2014) and mood instability (Eldar and Niv, 2015) and it has been suggested that the PE may play a role in choice itself (Dayan, 2009).

CHAPTER 3: Multiple time-linked reward signals in dACC

3.1. Introduction

The last chapter examined human behavior in a task in which the reward rate of a single option was changing systematically over time. In this chapter, fMRI data acquired during in the same experiment is analyzed using the parameterizations and the computational model described in the previous chapter (Wittmann et al., 2016a).

Previous studies have focused on decisions between two clearly defined and concrete options usually associated with specific stimuli. While this is clearly an important decision making mode, behavior often also reflects alternation between continuous engagement in the same behavior with behavioral change (Charnov, 1976; Hayden et al., 2011b; Neta et al., 2015; Stephens and Krebs, 1986). In such situations it is can be critical to know prospectively how profitable the next repetition of an action will be compared to previous ones.

Evidence comparison is central to decision making (Kira et al., 2015) and value comparison is central to reward-guided decision making. Most prominently, ventromedial prefrontal cortex (vmPFC) signals reflect the comparison of choice values when making decisions (Boorman et al., 2009). VmPFC blood oxygen level dependent (BOLD) activity is proportional to the difference in value between chosen and unchosen options and the chosen and unchosen options have dissociable positive and negative effects (Basten et al., 2010; Boorman et al., 2009; FitzGerald et al., 2009; Philiastides et al., 2010) which can be understood as reflecting a value competition process (Hunt et al., 2012; Jocham et al., 2012; Noonan et al., 2010; Strait et al., 2014).

A brain region comparing recent and past rates of reward requires fine-grained information about the past history of rewards. dACC is necessary for retaining information about the history of encountered rewards (Buckley et al., 2009; Kennerley et al., 2006).

Moreover, neurophysiological studies have demonstrated dissociable influences of past and recent rewards on the activity of dACC neurons (Bernacchia et al., 2011; Seo and Lee, 2007). In addition to dACC, activity in the dopaminergic system and striatum also reflects reward history (Bromberg-Martin et al., 2010; Hikosaka, 2007; Schultz, 2006). Reward history signals here have been understood in the context of reinforcement learning models (Sutton and Barto, 1998) that involve computation of a reward prediction error relative to previous outcomes (Bayer and Glimcher, 2005). However, typically, simple reinforcement learners cannot weight recent and past rewards in an opposing fashion. To do this, a learning mechanism requires a contrast mechanism comparing recent and past reward rates (Dayan, 2009; McNamara et al., 2013; Niv et al., 2007). Therefore, we used functional magnetic resonance imaging (fMRI) in humans to investigate value comparisons of recent and past reward rates.

3.2. Methods

3.2.1. *Imaging data acquisition and preprocessing*

Brain data were collected on a 3 Tesla Siemens Verio MRI scanner equipped with a 32 channel head coil. T1-weighted structural images were acquired using TR=3s, TE=4.75ms, TI = 1100ms, 1x1x1mm voxel size, 256x176x224 grid. Functional images were acquired using a Deichmann echo-planar imaging (EPI) sequence (Deichmann et al., 2003) with TR=3s, TE=30 ms, 3x3x3mm voxel size, 87° flip angle, 15° slice angle and z-shimming to reduce signal distortions and dropout in medial orbitofrontal brain areas.

Brain data were analyzed using FMRIB's Software Library (FSL) (Smith et al., 2004). FMRI data preprocessing for univariate analyses comprised spatial (Gaussian with full-width half maximum of 5 mm) and temporal filtering (3 dB cut-off at 100 s), motion correction using FSL's MCFLIRT and manual removal of noise components after visual inspection using FSL MELODIC. In a two-step procedure, preprocessed images were nonlinearly registered to Montreal Neurological Institute (MNI) space via subjects' structural images.

3.2.2. *FMRI whole-brain analyses*

FSL FEAT (Smith et al., 2004) was used for first level analyses. The fMRI time course was pre-whitened with FSL FILM to account for temporal autocorrelations. Motion regressors derived from MCFLIRT were included as nuisance regressors. Temporal derivatives of the regressors of interest were included and the model was temporally filtered before it was applied to the data. Group results on the second level were calculated with FSL FLAME 1 using outlier de-weighting and a cluster-forming threshold of $z>2.3$ and $p<0.05$. The only exception from this is presented in Fig.7b for illustration.

We ran three main fMRI designs; one based on the objective history of reward rates experienced at time of choice and the one derived from our RL-avgRR model. All designs modeled button presses as stick functions of no interest. The first (***fMRI GLM1;*** Fig.3a) contained a constant regressor time-locked to the leave-stay decision. This constant captured the duration of the last reward event (800ms) and the subsequent choice phase (2,000ms). We used lastRR, avgRR and the logarithm of the reaction times (logRT) as parametric modulators of the constant regressor. Based on the parameter estimates of lastRR and avgRR, we created the contrast lastRR-avgRR. We used this design to identify brain regions sensitive to the reward rate trend. ***fMRI GLM2*** (Fig.6a,b) was built identically, except for the parametric modulators of the LSD regressor. They comprised $PE_{expected}$(LSD-1) (abbreviated as just "$PE_{expected}$"), which was the expected prediction error, calculated with RL-avgRR, at the last time step before LSD. The only difference to the DV, i.e. to the expected prediction error at the time of the LSD ($PE_{expected}$(LSD); see equation 11 in last chapter) was that it had not yet been updated by the PE* of the last time step. In addition, it contained a binary choice regressor coding trials in which subjects decided to stay in a patch as "1" and trials in which they decided to leave as "-1", and logRT. Correlations between $PE_{expected}$(LSD-1) and choice are shown in Fig.2b for all subjects. ***fMRI GLM3*** (Fig.7) was also identical to the previous GLMs with the exception of the parametric regressors used. Again, we used $PE_{expected}$(LSD-1) and logRT. In addition we included the simple value estimate at the time of the last reward delivery and the magnitude of the last reward: value(LSD-1) and outcome(LSD-1). We calculated a standard prediction error on the contrast level by subtracting the value from the outcome regressor (Fig.7b). Note that for this GLM, RL-avgRR was fitted on the whole group instead of single subjects individually. The reason was that, although individual fits resulted in a set of regressors that were sufficiently decorrelated for the majority of

subjects, this was not the case for all of them. In these and all subsequent region of interest (ROI) analyses, all parametric and binary regressors were normalized.

3.2.3. *FMRI region of interest analyses*

ROIs had a radius of three voxels and were centered on peak voxels of significant clusters from fMRI GLM1 and fMRI GLM2 (Table 1). To guarantee statistical independence of ROI selection and ROI analyses, we used a leave-one-out procedure to identify ROI peak voxels for the analyses of main effects for areas identified in fMRI GLM 1. For this, we conducted group level analyses, leaving one subject out at a time. From the results of the remaining 19 subjects, we extracted local maxima of the relevant clusters and centered the ROIs for the left-out subject on the local maxima. We repeated this for all 20 subjects. Therefore, the ROI selection was statistically independent from the data of the subject that was subsequently analyzed in the ROI. For fMRI GLM2, no leave-one-out procedure was necessary because we did not test signals for significance in those ROIs that were related to the defining contrast. Note that correlation analyses were still conducted on the ROI centered on the peak of the whole group because those analyses are not affected by this potential problem.

For ROI time-course analyses, we extracted the pre-processed BOLD time courses from each ROI, averaged over all voxels of each volume. The time courses were normalized (per session, as for subsequent analyses), oversampled by a factor of 20 (using cubic spline interpolation, as for subsequent analyses) and, in a trial-wise manner, aligned at the time of LSDs. We then applied a GLM to each time point and computed one beta weight per time point, which resulted in a time course of beta weights for each regressor. We used two features of these beta weight time courses within an analysis window of four to thirteen seconds after LSD onset to investigate their relation to individual variance in behavior. First, to compute the slope of the beta-weight time-course signal, we identified

the subject specific absolute (positive or negative) peak of the beta weight time course within the analysis window and fitted a straight line on the time course from the time of LSD onset to the time of the absolute peak. Second, we averaged the time courses across subjects and identified the absolute beta weight group peak within the analysis window. We then took each subject's beta weight at that time point and examined its relation to behavior. FMRI GLM1 was calculated as a ROI time course analysis and resulting neural signals were correlated with lastRR and avgRR beta weights from behavioral GLM2 (see last chapter). fMRI GLM2 and fMRI GLM3, too, were calculated as a time course analysis and group peak signals were identified for significance testing (see below). Lastly, ***fMRI GLM4*** was analogous to behavioral GLM1 (see last chapter); it employed the reward rates in five discrete time bins relative to the leave-stay decision as well as logRT as parametric regressors. Group peak signals from this time course analysis are shown in Figure 4. Correlations with behavior were calculated using behavioral GLM1. fMRI GLM1 and fMRI GLM4 were similar in the sense that they parameterized objective reward history in complementary ways (Fig.1).

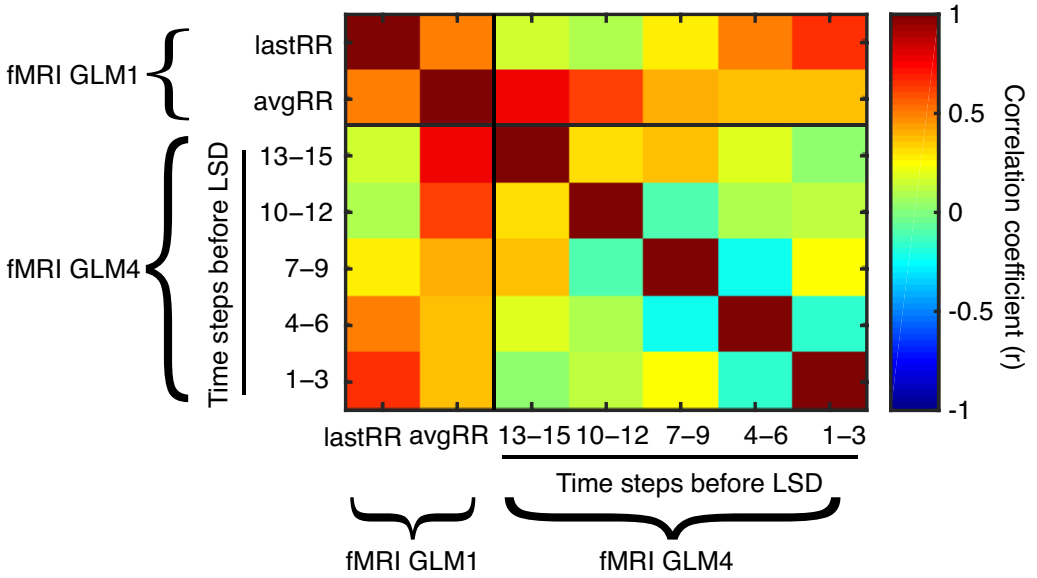


Figure 1. Correlation matrix of parameterizations of reward history used in two GLMS (fMRI GLM1 and fMRI GLM4). *The regressors used summarize reward history in complementary ways. Note that the parametric regressors within each GLM (lastRR and avgRR in fMRI GLM1 and the five time bins in fMRI GLM4) share less than 25% of their variance.*

We performed significance testing on time course analyses using a leave-one-out procedure on the group peak signal (Fig.6c,d and Fig.7a) to avoid potential temporal selection biases. For every subject, we calculated the time course of the group mean beta weights of the relevant regressor based on the remaining 19 subjects. We then identified the (positive or negative) group peak of the regressor of interest within the analysis window of four to thirteen seconds from LSD onset. Then, we took the beta weight of the remaining subject at the time of the group peak. We repeated this for all subjects. Therefore, the resulting 20 "peak" beta weights were selected independently from the time course of the subject analyzed. We assessed significance using t-tests on the resulting beta weights.

3.2.4. *Supplementary fMRI analyses*

We ran a couple of control analyses, which did not pursue our main line of investigation but instead aimed at distinguishing our effects of interest from brain activity related to other factors, in particular task difficulty and choice.

We constructed a measure of response selection or task difficulty (Kolling et al., 2016). We did this to clearly dissociate our neural effects of interest from difficulty effects that for example a conflict account of dACC might predict (Botvinick et al., 1999). We used behavioral GLM4, which was used to predict subjects' choices with the two regressors lastRR and lastRR-avgRR (see last chapter). For every subject, we combined individual beta weights of lastRR and lastRR-avgRR with the parameter values of lastRR and lastRR-avgRR in a logistic GLM to predict the stay probabilities on every trial. We also included the subject specific constant. For all subjects, we derived a measure of task difficulty for each trial:

$$\text{Task difficulty} = -\,|\text{stay probability} - 0.5|$$

Task difficulty was not systematically correlated with our regressors of interest, neither for our RL-free analysis approach using lastRR and avgRR (Fig.2a), nor for our RL-avgRR based analysis approach (Fig.2b).

For our control GLMs, we included the task difficulty regressor as an additional parametric contrast time-locked to the decision phase of the experiment (as the other parametric regressors) in *fMRI GLM1* and *fMRI GLM2*. All other aspects of these GLMs remained unchanged. Moreover, we also reran *fMRI GLM4* with the inclusion of a binary choice regressor (described above) time-locked to the decision phase of the experiment (as the other parametric regressors). Again, all other aspects of the GLM remained unchanged.

Finally, as a first approximation of our RL-avgRR based analysis, we constructed a GLM (***fMRI GLM5)*** using the same constant regressors (i.e. the regressors indicating

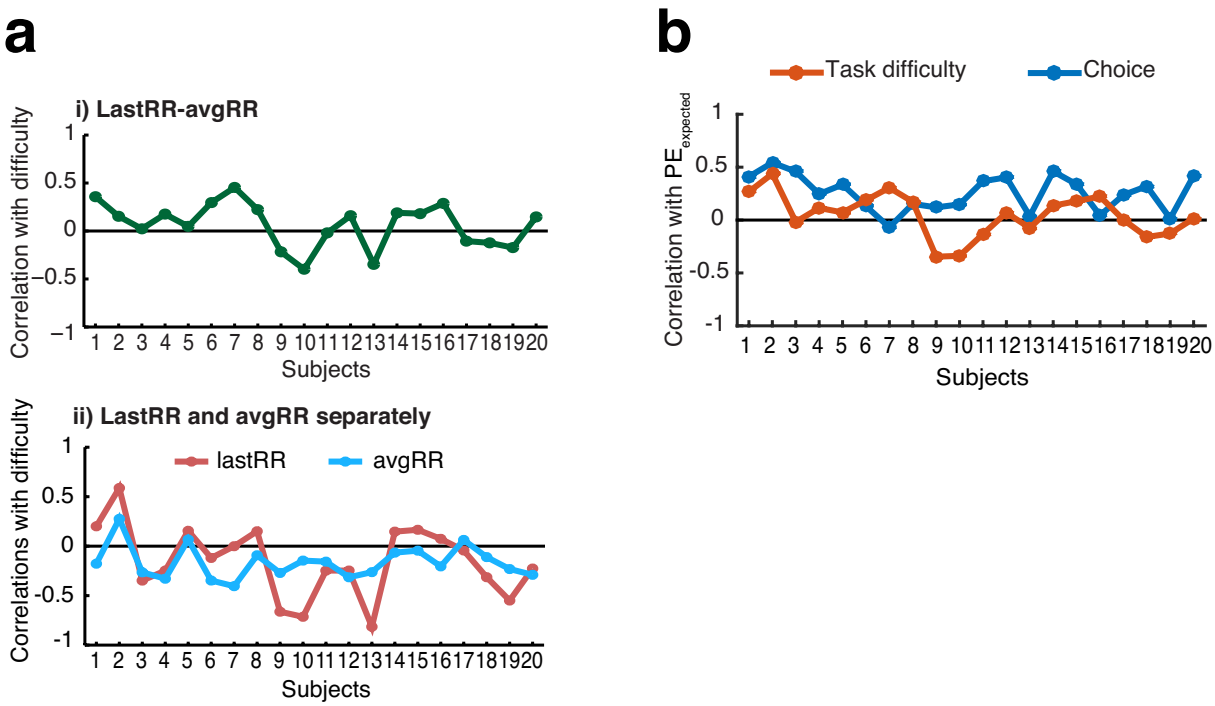


Figure 2. Task difficulty was not systematically related to choice variables. *(**a**) Task difficulty did not correlate in a systematic way with lastRR-avgRR (i). The same was true for its correlations with the components lastRR and avgRR (ii). (**b**) The same the case for our RL based analysis approach. The expected prediction error one time point before choice ($PE_{expected}$(LSD-1)), abbreviated as just " $PE_{expected}$", showed no strong or systematic correlation with task difficulty. In addition, correlations of $PE_{expected}$ with a binary choice regressor are shown.*

occurrence of key trial events such as LSD time) as in fMRI GLM 1,2 and 3. However, for this GLM, the parametric regressors time-locked to the LSD comprised only logRT and the decision variable (DV) of the RL-avgRR model, which was the value estimate of the higher-order part of RL-avgRR model at the time of the LSD ($PE_{expected}$ (LSD)).

3.3. Results

3.3.1. *dACC contrasts recent and past rewards*

We began our fMRI analysis by identifying brain regions involved in the computation of the reward rate trend using a parametric approach that does not rely on an RL model. We used the regressor of lastRR-avgRR, which reflects the reward rate trend. Activity in dACC, right ventromedial striatum, and right frontal operculum (FO) increased at the time of the LSD when the reward rate trend was more positive, leading subjects to stay in an environment (Fig.3a; Table 1). We went on to conduct regions of interest (ROIs) analyses using a leave-one-out procedure to identify individual ROIs (Methods). All ROI analyses were time-locked to the LSD.

To investigate whether reward rate trend effects in dACC were, as hypothesized, composed of opposing signals indexing recent and past reward rates, we obtained separate estimates of activity related to the regressors lastRR and avgRR in the same GLM. In dACC (Fig.3b), beta weights indicated a positive effect of lastRR (t_{19}=2.78; p=0.012) and a negative effect of avgRR (t_{19}=-3.8; p=0.001), recalling the positive and negative effects of these variables on behavior (see last chapter). To examine the relationship between individual differences in neural effects and behavioral effects, we estimated slopes of signal increase as indices of evidence accumulation (Jocham et al., 2012; Kolling et al., 2012) in dACC. As previously, we used slopes of signal increase because they are simply and intuitively related to evidence accumulation processes that may occur in dACC (Hayden et al., 2011b; Kolling et al., 2012) although it is possible that other aspects of dACC BOLD are also related to behavior. The dACC avgRR slope in each subject correlated with the weight each subject placed on the same information when making a decision (r=0.61; p=0.004). A similar relationship between dACC lastRR slope and

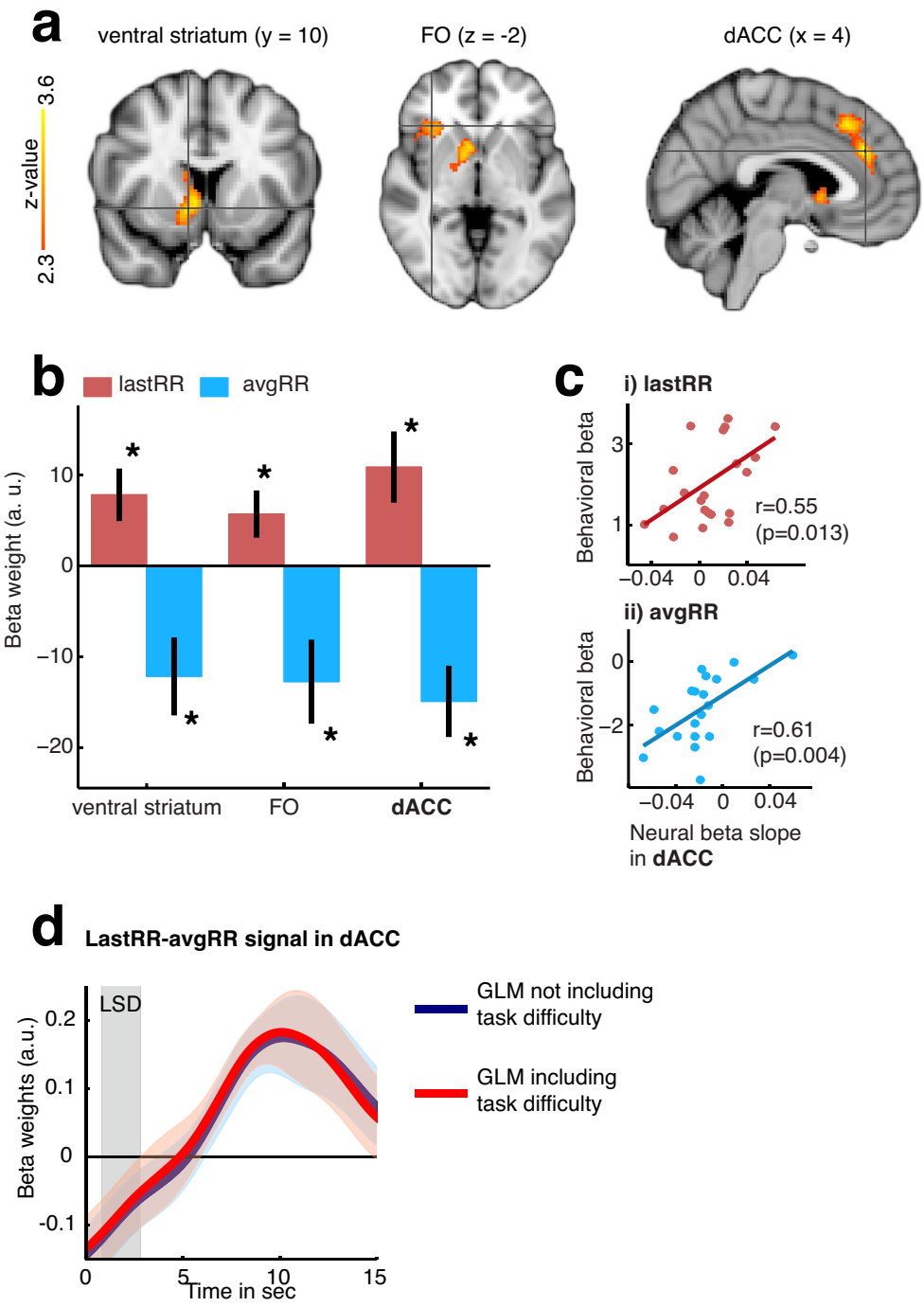


Figure 3. Opposing effects of recent and past reward rates in dACC predict choice - 1. *(**a**) A whole-brain contrast lastRR-avgRR time-locked to the leave-stay decision revealed three areas in which a more positive reward rate trend led to more activity: dorsal anterior cingulate cortex (dACC), right frontal operculum (FO), and right ventral striatum. (Family-wise error cluster-corrected, z>2.3, p<0.05). (**b**) ROI analyses of the three areas (using leave-one-out procedures) show separate neural responses to lastRR and avgRR. For all areas, extracted beta weights indicate that lastRR had a positive effect, while avgRR had a negative effect. (**c**) In dACC, the slopes of the behavioral beta weights for both lastRR (i) and avgRR (ii) were predictive of how much the respective recent and past reward rates influenced subjects' choices. Neither of the other areas showed either correlation. (**d**) fMRI time course of lastRR-avgRR effect in dACC (blue line). In a control GLM, we included a task difficulty regressor in addition, but the effect of lastRR-avgRR in dACC remained unchanged (red line). (Error bars are s. e. m.; *, p<0.05).*

lastRR behavioral weight confirmed the relationship (r=0.55; p=0.013) (Fig.3c). Similar analyses revealed that although lastRR and avgRR were associated with opposing effects in FO and ventral striatum (all t_{19}>|2.2|; all p<0.041) there were no correlations between neural and behavioral effects (all p>0.05).

Moreover, inclusion of response selection difficulty (Kolling et al., 2016) in our GLM did not change the lastRR-avgRR signal in dACC (Fig.3d). The lastRR-avgRR effect after including task difficulty showed a slight, but not significant increase in peak effect size relative to the lastRR-avgRR effect from the original GLM (*fMRI GLM1*). This is consistent with previous findings that foraging-related decision variables were encoded in dACC beyond an unspecific effect of task difficulty in this brain region (Kolling et al., 2016; Kolling et al., 2012).

To validate our first analysis, we used a complementary approach to investigate the reward effects in dACC in a temporally more detailed manner. We binned rewards in five time steps as in our behavioral analysis (see last chapter; binned in sets of three time steps: LSD-1-3, LSD-4-6, etc.). This allowed investigation of the parametric effects of rewards encountered at multiple time points in the past on present BOLD activity. We extracted the group peak signals for each time bin using a time course analysis (Methods). For dACC neural results were, again, strikingly similar to behavioral results (Fig.4a-i): a gradual transition from positive to negative effect sizes was observed as reward rates became more distant in time. The same was true for FO but not the ventral striatum (Fig.4a). To quantify similarity between the three areas, we calculated a 3 (ROI) x 5 (time bin) ANOVA and found a significant effect of time bin ($F_{2,47}$=8.39; p=3.3*10^{-4}) and an interaction effect ($F_{8,152}$=3.68; p=0.001) suggesting differing temporal gradients of neural response to reward rates (additional analysis of another ventrostriatal ROI also failed to identify monotonic changes in effects of the sort seen in ACC and FO).

In the behavioral analysis, changes in beta weights reflect how rapidly the impact of reward rates on the tendency to stay in a patch reversed with time. For each subject, we fitted a regression line through the five data points from both the neural analysis described here and the corresponding behavioral analysis and took the slopes of these lines as indices of temporal gradients of neural and behavioral effects (Fig.4b-i). Given the close relationship of neural activity in dACC to behavior found in the previous analysis (Fig.3c), we tested whether the temporal gradient of neural activity in dACC was predictive of the temporal gradient of reward effects on choice. Again we found a significant correlation for dACC but not the two other regions (r=0.52; p=0.018; Fig.4b-ii).

Visual inspection of the neural beta weights for all reward bins Figure 4a may suggest that the difference between cortical areas and the ventral striatum only relates to the earliest time bin LSD-13-15. To show that the significant correlation between neural and behavioral gradient in dACC (Fig.4b-ii) is not only dependent on this earliest time bin, we repeated the same analysis, but removed the earliest time bin from the calculation of the neural and behavioral temporal gradient. The correlation remained significant (r=0.49; p=0.027; Fig.4b-iii).

In another control analysis, we recalculated the time bin GLM for dACC shown in Figure 4a, but included in addition a regressor controlling for choices subjects made. This binary choice regressor contained "1" for stay and "-1" for leave choices. Fig.4c shows the neural signals of the five time bins when including such a regressor. Visual inspection shows only minimal change between both GLMs. Accordingly, in a 2 (GLM) x 5 (Time bin) ANOVA, we found differences in neural response to the time bins (main effect time bin: $F_{4,76}$=4.6; p=0.002) but no difference between both GLMs (main effect GLM: $F_{1,19}$=2.18; p=0.156) and no interaction effect ($F_{4,76}$=1.12; p=0.318).

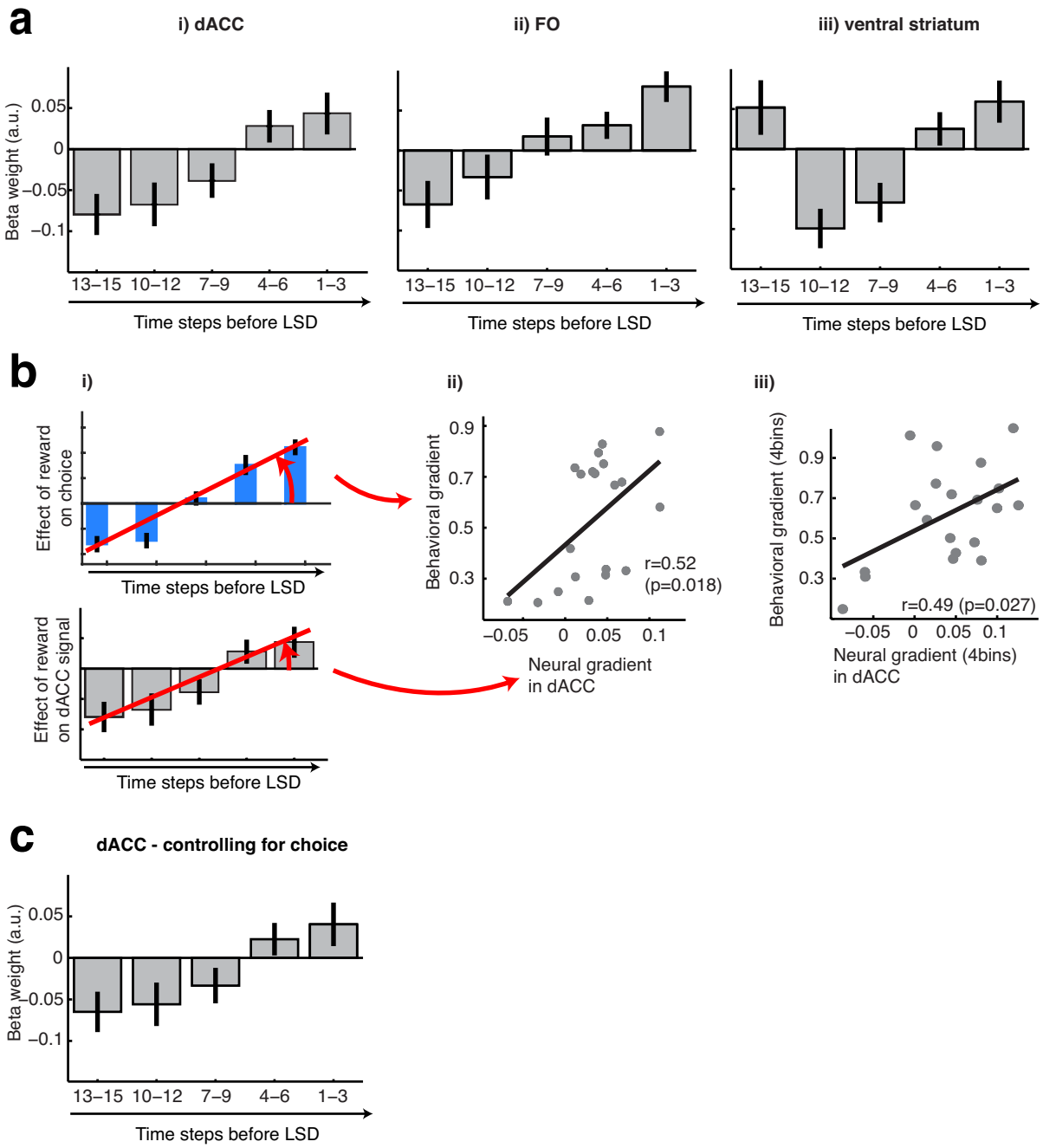


Figure 4. Opposing effects of recent and past reward rates in dACC predict choice - 2. *(**a**) In dACC, we validated the results found in Figure 3 by analyzing the neural effects of reward rates presented in discrete time bins before the leave-stay decisions (analogous to the behavioral analysis in the last chapter). We found a temporal gradient of reward effects on BOLD activity that was similar to the temporal gradient of reward effects on behavior. (**b**) Illustration of the gradients in effect sizes in behavior (i, top) and brain signals (i, bottom), expressed by the slope of the red line. Using this alternative analysis approach we were again able to confirm a relationship between dACC activity and behavior; in dACC, but in neither of the other areas, the gradient of neural responses to past rewards was predictive of the behavioral gradient characterizing the influence of past rewards on the decision to stay or to leave (ii). This correlation remained significant also when only considering the most recent four time bins (iii). (**c**) dACC effect sizes for the five time bins when controlling for general choice effects. (Error bars are s. e. m.).*

In sum, regardless of the precise manner in which BOLD signals were decomposed, dACC signals predicted the positive effects of recent and the negative effects of past rewards on choices. The opposing time-linked signals observed do not suggest that dACC and the other regions integrate rewards to a simple mean estimate (as RL-simple would), but instead point towards a comparison of recent and past reward rates.

3.3.2. *Prediction errors and choice signals in medial prefrontal cortex*

We next examined the neural mechanisms of reward trend guided choice based on the reinforcement learning model that captured subjects' choices best: RL-avgRR. For each subject, we derived trial-by-trial parameters from an individually fitted RL-avgRR model. First, we used the decision variable of RL-avgRR, the expected prediction error at the time of the LSD, in a whole-brain GLM time-locked to the LSD. The expected prediction error allows subjects to make inferences about the reward likely to be received in this environment over the 15 time steps that might follow the LSD. We again identified an activation cluster in dACC (Fig.5; Table 1), but neither in the ventral striatum nor the frontal operculum. This result is well in line with the previous finding that dACC signals reflect previous reward in a choice related way (Fig.3,4). In addition, the contrast identified an area around the right inferior frontal junction (IFJ)(Neubert et al., 2014).

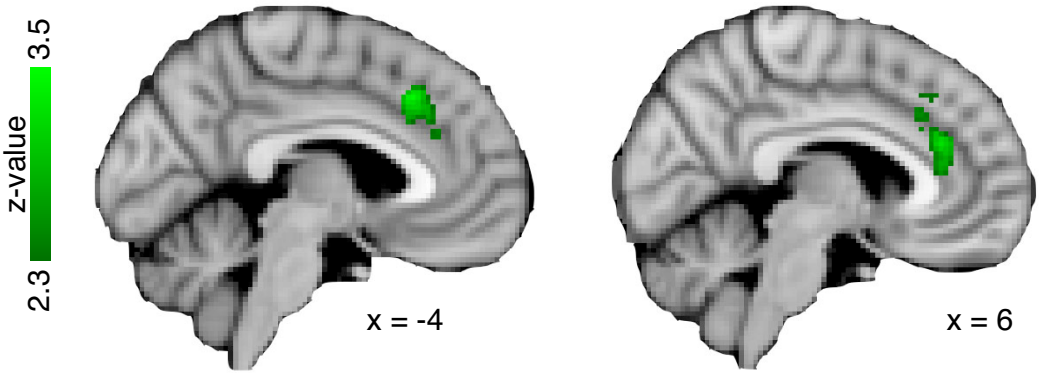


Figure 5. Representation of RL-avgRR's decision variable in cingulate cortex.

We went on to investigate the relationship between expected prediction errors, choice, and instantaneously experienced prediction errors in two analyses. Both analyses were time-locked to the LSD and used the value of the expected prediction error at the time step of the last reward delivery, i.e. one time step before LSD, not yet updated by the last reward event: $PE_{expected}$(LSD-1). For simplicity, we refer to this parameter just as $PE_{expected}$ in the remaining section.

In a first GLM, we aimed to dissociate neural representations of the reward trend-related decision variables from the decisions they drive. We used a binary choice regressor and $PE_{expected}$, which were sufficiently decorrelated so that both could be included in the same GLM (Fig.2b). Consistent with previous analyses, we found strong effects of $PE_{expected}$ in several brain regions (Table 1) including dACC (Fig.6a) but excluding frontal operculum or ventral striatum. Although the dACC activation was more posterior than previously it remained within the anterior rostral cingulate zone (Neubert et al., 2015). In order to understand the function of adjacent component regions with dACC regions we examined activity in ROIs placed over the dACC peak we identified in this analysis (*posterior dACC*; red in inset panel, Fig.6c,d) and the closely adjacent dACC peak we identified in the first fMRI analysis in Fig.3a (*dACC*; yellow in inset panel, Fig.6c,d). Note that we used a leave-one out procedure to select the *dACC* ROI. We also used a leave-one-out procedure to avoid temporal bias in the signal analysis (Methods). By contrast, vmPFC was more active for stay compared to leave choices (Fig.6b) consistent with reports that vmPFC activity reflects choice computations in a variety of contexts (Chau et al., 2014; FitzGerald et al., 2009; Hunt et al., 2012; McGuire and Kable, 2015; Plassmann et al., 2007).

Further ROI analyses, using similar leave-one-out procedures to avoid bias, suggested vmPFC only carried information about the ultimate stay/leave decision and not about the expected PEs on which the decision was based, while *posterior dACC* (Fig.6a)

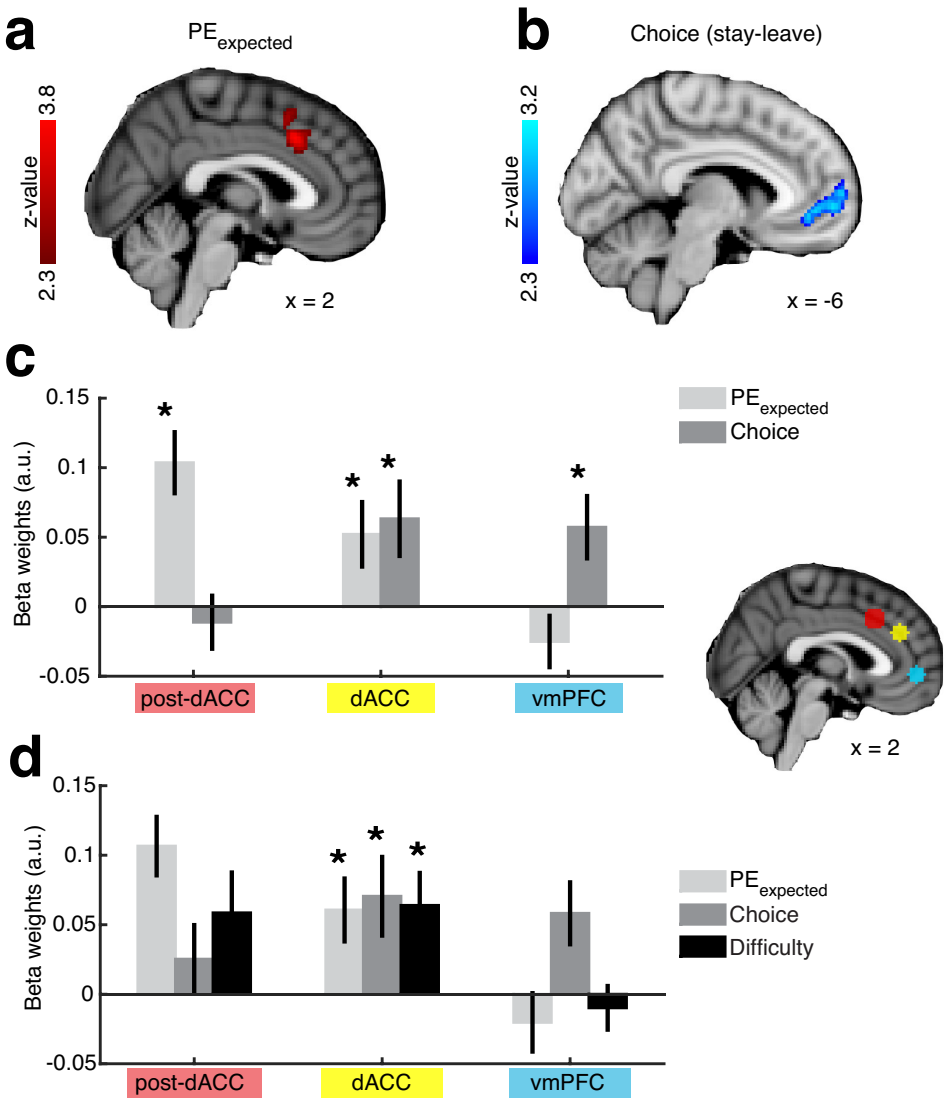


Figure 6. Separable representations of choice and decision variables along the ACC. *(**a**) At the time of choice, activity in a posterior region of the dorsal anterior cingulate cortex (post-dACC) varied as a function of expected prediction error ($PE_{expected}$). Activity was largely confined to the anterior rostral cingulate zone. The same was the case previously for dACC activity related to the reward rate trend (Fig.3a). (**b**) The vmPFC BOLD signal increased when subjects decided to stay in a patch compared to leaving it. (**c**) Analysis of BOLD response to the model based evidence for staying in a patch ($PE_{expected}$) and the ensuing choice along an axis of post-dACC (red, from panel a), a more anterior dACC region (yellow, lastRR-avgRR contrast from Fig.3a) and vmPFC (blue, from panel b). Anterior dACC signals encode both $PE_{expected}$ and the ensuing choice to commit to or leave the patch. Contrarily, post-dACC and vmPFC show only a significant effect of $PE_{expected}$ and the categorical choice, respectively. Note that the anterior dACC ROI was identified using a leave-one-out procedure. Significance of $PE_{expected}$ in post-dACC and choice in vmPFC was assessed in the previous whole-brain analysis (panels a,b). (**d**) Control analysis including task difficulty regressor. Note that the $PE_{expected}$ effect in post-dACC and the choice effect in vmPFC were both significant in the initial whole brain analysis that was used to first identify the ROIs but here, to avoid double-dipping, the $PE_{expected}$ effect in post-dACC and the choice effect in vmPFC, are only shown for illustration and were not further tested for significance (it is for this reason that we did not add asterisks for the pale grey bar for post-dACC and the dark grey bar for vmPFC). (*, $p<0.05$; error bars are s. e. m. between subjects).*

only carried information about the expected PEs but not the ultimate leave/stay decision. By contrast, the *dACC* exhibited significant effects of both expected PEs (t_{19}=2.26; p=0.037) and choice (t_{19}=2.37; p=0.029; Fig.6c). This replicates our previous result (Fig.3,4) suggesting the presence of both reward and choice related effects in *dACC*.

To consolidate our results, we repeated this analysis and added our previous task difficulty regressor (Kolling et al., 2016). As in our previous control analysis (Fig.3d) reward signals ($PE_{expected}$) and choice signals in dACC remained significant and even exhibited very small increases in effect sizes (t_{19}=2.68; p=0.015 and t_{19}=2.49; p=0.022, respectively; Fig.6d). In addition and not unexpected for a brain region that computes choice variables, task difficulty had a significant effect in *dACC* (t_{19}=2.74; p=0.013) but not in the other two regions (*posterior dACC*: t_{19}=2.02; p=0.058; vmPFC: t_{19}=-0.62; p=0.54).

In a second GLM, we aimed to identify neural correlates of the component parts of RL-avgRR. As regressors, we used the three component variables that the model needs to track at the time point before choice (t=LSD-1): $PE_{expected}$, value, and outcome. Value and outcome refer to the model's simple value estimate and to the magnitude of the last reward event, respectively (see last chapter for a mathematical definition). In particular, we investigated whether brain regions encoded a standard prediction error, i.e. increased their activity as a function of outcome, but decreased their activity as a function of value (contrast outcome minus value) (Behrens et al., 2008; Schultz, 2006). Note that for this analysis, RL-avgRR was fitted on all subjects to ensure that, for all subjects, the three regressors of interest shared less than 25% of their variance.

Again, $PE_{expected}$ identified a cluster in dACC, amongst other regions (Table 1). Based on previous reports (Daw et al., 2011), we strongly expected the ventral striatum to encode standard reward prediction errors. To test this, we used the ventral striatal ROI

from our previous analyses (Fig.3b), which was derived using a leave-one-out procedure, as well as the same striatal ROI mirrored to the contralateral side (individually for each subject to avoid bias). Indeed, we found significant PE signals (contrast outcome minus value) bilaterally in ventral striatum (left: t_{19}=3.31; p=0.004; right: t_{19}=3.04; p=0.007; Fig.7a). This PE effect did not survive standard whole-brain cluster correction. However, when lowering the cluster-forming threshold (|z|>2.3, p>0.15), the first PE signal to reach whole brain significance was centered on the left ventral striatum (MNI peak at: -6/12-10; Fig.7b).

In sum, we found that dACC most reliably encoded the expected prediction error. Although the precise clusters identified using the expected prediction regressor varied as we changed the time point to which the regressor related (LSD or LSD-1) or whether the

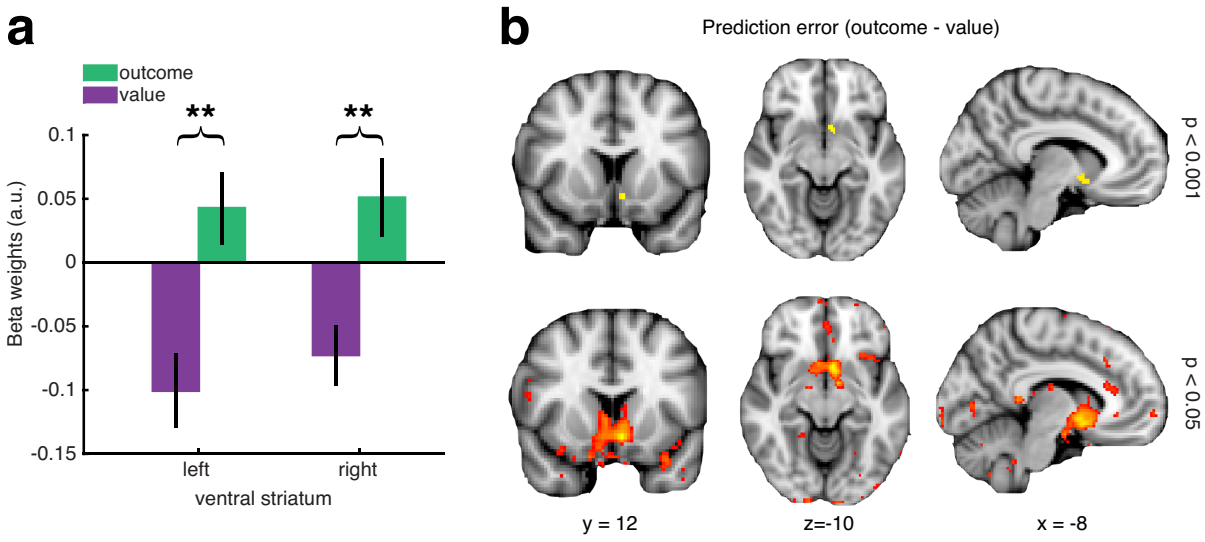


Figure 7. Standard prediction errors in the ventral striatum. *(**a**) Significant standard prediction error (outcome minus value) signals in left and right ventral striatum. Regions of interest are the same as for the striatal ROI in Fig.3b as well as the same ROI mirrored to the contralateral side. For each subject and both hemispheres, ROI were determined via a leave-one-out procedure to avoid spatial bias. (**b**) Whole-brain prediction error contrast shown at two thresholds, shown for illustration only. The prediction error signal is centered on the left ventral striatum (top row; threshold of 0.001, uncorrected). No other brain region showed an equally strong encoding of standard prediction errors (bottom row; threshold of 0.05, uncorrected). Note that images are shown according to the radiological convention, so left/right is flipped. X/y/z coordinates refer to both images from both rows. (**, p<0.01; error bars are s. e. m. between subjects).*

RL-avgRR was fitted individually or on the whole group, dACC was the only region that was identified in all these analyses as well as in the previous analysis that did not employ any RL model (Fig.3,4). Moreover, the ventral striatum was most sensitive to instantaneous standard prediction errors and vmPFC showed signs of choice-related activity in the absence of a specific representation of the reinforcement history.

Contrast	Region	Peak Coordinates x/y/z (in mm MNI space)	Z-Value
last-avgRR (fMRI GLM 1)	Dorsal anterior cingulate cortex (dACC; RCZa)	6 38 28	3.93
	Right frontal operculum (FO)	34 26 -2	3.08
	Right ventral striatum	10 10 -4	3.49
Choice (fMRI GLM2)	ventromedial prefrontal cortex (vmPFC)	-2 50 -2	3.19
$PE_{expected}$(LSD-1) (fMRI GLM2)	posterior dorsal anterior cingulate cortex (post-dACC; RCZa)	2 20 38	3.82
	right inferior frontal sulcus (IFS)	38 28 22	3.82
	right inferior parietal lobule	56 -38 44	3.63
	right Thalamus	8 -10 8	3.52
$PE_{expected}$(LSD-1) (fMRI GLM3)	right inferior frontal gyrus	44 10 24	3.68
	Dorsal anterior cingulate cortex (RCZa)	-4 22 40	4.25
	right ventral striatum	8 6 0	3.65
	right frontal operculum	30 22 -4	3.66
	left frontal operculum	-32 18 -6	3.49
$PE_{expected}$(LSD) (fMRI GLM5)	Right inferior frontal junction (IFJ)	52 10 26	3.52
	anterior rostral cingulate zone (aRCZ)	-6 24 40	3.46

Table 1. Peak coordinates of significant clusters in fMRI contrasts (|z|>2,3, p<0.05, cluster corrected).

3.4. Discussion

Considerable emphasis has been placed on the way both human and animal behavior is guided by value expectations based on learned associations between specific stimuli and reward outcomes. In many natural foraging situations, however, part of what animals are doing may simply be deciding whether to continue a behavior or whether it will be more profitable to change behavior (Charnov, 1976; Hayden et al., 2011b). In such situations a comparison of the current shorter-term rate of reward with the longer-term average reward rate of the environment is highly informative. The opposing, time-linked reward representations in dACC (Fig.3,4) allow switching behavior as a function of the changing reward rates.

We found that activity in several brain areas including dACC, FO, and striatum reflects multiple time-linked reward representations. The signs of the brain signals reflected the opposing weights of rewards in recent and distant past on choice (Fig.3b,4a) but only in dACC were these weights predictive of the weights each subject placed on recent and past rewards in their behavior (Fig.3c,4b). Such relationships between dACC activity and reward trend-guided decisions are analogous to those found between individual variation in vmPFC signals and individual variation in value-guided binary and trinary choice behavior (Boorman et al., 2009; Chau et al., 2014; Strait et al., 2014). In addition, a role for dACC in deciding based on extrapolated reward trends or expected future rewards is consistent with a recent proposal that ACC is involved in evaluating future strategies before they are being executed (Schuck et al., 2015). Notably neurons in macaque dACC and interconnected areas represent memories of reward rates with different time constants (Bernacchia et al., 2011; Bromberg-Martin et al., 2010; Nakamura et al., 2012; Seo and Lee, 2007). The behavioral role of such activity patterns has not been clear. Different reward memories could be used to adapt the learning rate of an agent to the

volatility of an environment (Behrens et al., 2007b). Another possibility is neural populations with different learning rates interacting with each other might compute a reward rate trend. It is possible that the neurons with multiple time constants described in dACC in macaques may correspond to just such a system (Bernacchia et al., 2011).

Although there might be other models that capture reward trend guided choice, using the expected prediction error to solve this problem is consistent with several behavioral (see last chapter) and neural findings. The brain regions that carried expected prediction error signals in our study, in particular dACC, have been linked to lower and higher order prediction errors before. For instance, activity of neurons identified recently by Bernacchia and colleagues (Bernacchia et al., 2011) (e.g. their Fig.3d,h,l) are influenced by past reward outcomes in a very similar way as the expected prediction errors are derived from past reward outcomes in our model (Fig.6a-iii in the last chapter). These neurons, which have been found in several brain regions including anterior cingulate cortex, are characterized by a double exponential function in which recent rewards have a sign reversed influence on activity compared to rewards in the more distant past, mirroring the behavioral and neural effects of past rewards we have found. Finally, expected prediction errors are also consistent with recent suggestions that dACC plays a role in predicting the prediction errors of lower computational layers (Alexander and Brown, 2015).

In agreement with the first neural analyses (Fig.3,4), using the modified RL model, RL-avgRR, we were again able to show that reward trend related representations, in this case the expected prediction error, are linked to dACC (Fig.5,6). The ventral striatum, by contrast, signaled standard prediction errors (Bayer and Glimcher, 2005; Daw et al., 2011; Schultz, 2006) i.e. the difference of outcome and value estimate (Fig.7). Such a standard PE signal may feed in a cortical network, arguably comprising dACC, which computes a longer-term estimate of prediction errors. Moreover, using the modified RL model, we

were able to dissociate the expected prediction errors and representation of the choices that subjects ultimately made (Fig.6c). We found a graded transition from representation of expected prediction errors to choice moving from posterior to more anterior dACC and on into vmPFC, reminiscent of similar functional gradients in prefrontal cortex (Badre et al., 2009; Koechlin et al., 2003). The finding that dACC activity was related to both reinforcement-history-derived variables and ensuing choices further supports the contention that this area represents and translates reward history representation into a choice-related representation, as had also been suggested by the correlation between dACC neural and behavioral beta weights (Fig.3c,4b).

Notably choice-related signals in vmPFC and dACC were aligned in this experiment; both areas are more active for stay compared to leave choices and dACC activity increases with evidence for staying. In previous experiments vmPFC activity was positively related to the value of staying versus discarding a current choice, while dACC was positively related to exploring versus exploiting a choice (Daw et al., 2006; Doll et al., 2015; Kolling et al., 2012; Kolling et al., 2014; McGuire and Kable, 2015). However, while the value of the stay-option in the previous experiments was relatively well known and stable, it had to be inferred in the current experiment, because the future value of a patch is, by design, different from its past value. The leave-option was, on the other hand, pre-learned and stable. In other words: the outcome uncertainty for the stay-option is, unusually, higher than for the leave-option. Therefore, the present results can be integrated with previous findings by suggesting dACC is coding value in a framework tied to pursuing the more uncertain option; by contrast vmPFC may code the value of staying with the current or default choice (Boorman et al., 2013b; Kolling et al., 2012; McGuire and Kable, 2015; Scholl et al., 2015). The two reference frames are aligned in the current experiment but have been opposed in previous experiments.

CHAPTER 4: Predicting actions of oneself and others

4.1. Introduction

Social environments require humans and other primates to monitor others (Chang et al., 2013a; Chang et al., 2013b) and to know not only their own abilities but also the abilities of others. This knowledge guides establishment of social dominance hierarchies (Zink et al., 2008) and can be linked to features of brain structure and function (Noonan et al., 2014; Sallet et al., 2011). It guides animals' choices in a powerful way. For example, the decision to engage in fundamental modes of social interaction such as cooperation or competition with a conspecific is guided by knowledge about their abilities and social status relative to one's own (Wang et al., 2011). Estimating abilities of both self and others on the basis of past performance may be particularly important for humans, as they are able to coordinate and execute multi-step tasks such as building a shelter together with others (Misyak et al., 2014; Tomasello et al., 2005) or making complex decisions in groups (Kerr and Tindale, 2004). Moreover, they have to use this knowledge flexibly because social alliances can change very quickly: a competitor within one's own company might quickly become a cooperator when competing with a different company.

It has been a long-standing idea in psychology that humans derive expectations about whether they will succeed in a given task from their past task performance (Bandura, 1977). At the same time, people do not learn about themselves in isolation but relative to their social environment. Comparisons with other people can be used as an effective means for self-evaluation (Festinger, 1954; Mussweiler, 2003) and, conversely, people base judgments of other people on knowledge of their own traits (Allport, 1924; Krueger, 2007). However, the social influence on judgments of self and others can vary. We might be influenced more strongly by others simply because we like them (Heider, 1958); we might, by default, evaluate members of our own group more positively than members of a

different group (Brewer, 1979), and we might perceive others as more similar to us when we cooperate with them than when we compete with them (Toma et al., 2010). In sum, learning about self and others is often based on all three aspects: objective experience, self/other comparisons, and the social context.

Here we test, first, whether we estimate our abilities from monitoring our performance over time just as we estimate the values of our actions from monitoring their outcomes (Daw et al., 2011). We show that the history of an individual's performance is indeed used to estimate their ability. Second, we test if ability estimated for oneself is also dependent on the performance of other people. We find that, surprisingly, the performances of other people also influence individuals' assessments of their own abilities. When we are cooperating with someone who performs well our estimates of our own ability are inflated, and when we are cooperating with someone who performs poorly our ability estimates are depressed. Reciprocally, the ability of another individual is not only estimated from the other's performance but also from one's own performance. We demonstrate that such self-other-mergence not only impacts on people's subjective evaluation of themselves and others, but even affects how they subsequently adjust their performance. We refer to ability estimation for self and others based on their respective performance history as "appropriate", while we refer to the misattribution of past self-related performance when estimating others' abilities (or other-related performance to estimate one's own ability) as self-other-mergence (SOM).

This chapter introduces a novel experimental paradigm that was used to investigate self and other performance monitoring. Human behavior in that task was analyzed using general linear models (GLMs) and computational modeling (Wittmann et al., 2016b). The subsequent chapter presents fMRI data acquired during the same experiment.

4.2. Methods

4.2.1. *Subjects*

Twenty-six subjects participated in the experiment. Two were excluded from data analysis due to excessive motion (final sample: 24 subjects; nine female; aged 19–31). All provided informed consent. The study was approved by the ethics committee of the University of Oxford (MSD-IDREC-C1-2013-133). Subjects received £35 as a show-up fee and a bonus based on task performance (range: £4.50 to £11.80).

4.2.2. *Overview over experiment structure*

On each trial subjects performed a reaction-time minigame. They were told that two other players independently performed the minigame at the same time. We explain the nature of the minigames in detail below (see also Fig.9). The minigames' precise nature is less critical than the fact that they provided a vehicle to investigate how subjects developed an estimate of their ability that was based on their performance and how this changed

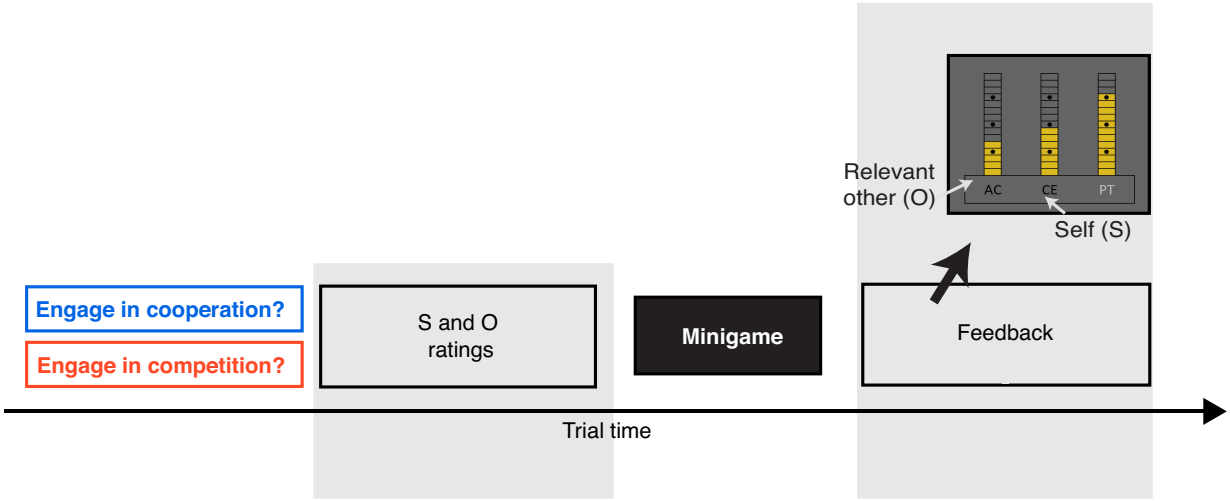


Figure 1. Schematic of trial timeline. *On every trial, all players played a minigame, after which parametric performance feedback was provided (upper right; higher bars indicate better performance) that enabled performance to be learned. Letters indicate subject initials for Self (S; middle position) and two other players. A relevant other (O) was pre-specified on each trial but changed between trials. Ratings and minigames were embedded in engage/avoid decisions that established either a cooperative or competitive context.*

depending on interactions with two other players whose performances they also saw. On each trial in the experiment subjects performed a short trial of a minigame and parametric feedback about their own performance and the performance of the other players was provided at the end of each trial (Fig.1). Subjects could use the performance feedback to form ability estimates for self and the two others over the course of the experiment.

We used pre-determined performance feedback schedules to carefully match observed performance for self and others and to keep them stable across subjects. This ensured that performance learning for self and others were comparable and that individual differences in task behavior were interpretable. Subjects were told that the performance feedback reflected their objective performance mapped on a 15-point performance scale and that the previously established mapping was the same for all players. Therefore, subjects received explicit and independent performance feedback for all players. Using not only one but two minigames ("time task" and "color task") in pseudorandom trial order made it possible to have, on the one hand, slowly drifting performance shifts within a minigame (as abilities are thought to be relatively stable features (Boorman et al., 2013a)) but on the other hand, reduced sequential correlations across trials (by switching between

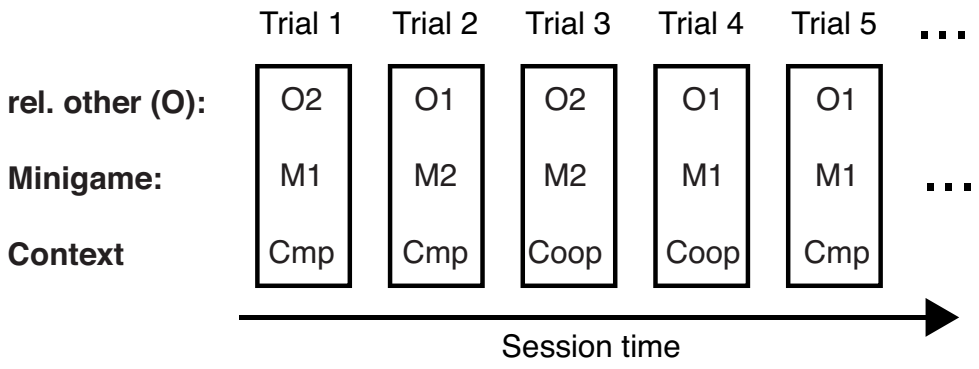


Figure 2. Schematic of session timeline. *Each trial is characterized by one of two other players having the role of relevant other (O1/O2), one of two possible minigames (M1/M2) and has either a cooperative (Coop) or compete (Cmp) social context. All three trial features were experimenter determined and pseudorandomly interleaved.*

minigames that were performed at different levels; Fig.2) and ensured a full parametric range of performance feedback thereby making it possible to perform event-related fMRI analysis. Finally, having two minigames allowed us to establish the generality of our findings.

Each trial was performed either in a cooperative or a competitive context. For example, on some trials subjects were given the opportunity to cooperate with one of two other players in the next run of the minigame (Fig.3a). This meant that their own performance score and the other player's performance score would be summed together and if it exceeded a threshold indicated on the screen (varying trialwise) then points were awarded (which were translated into monetary reward at the end of the experiment). Subjects could first decide whether to engage in cooperation or whether to avoid it (engage/avoid decision). If they avoided cooperation then the points awarded were randomly distributed around zero. The other player was pre-specified on each trial and is referred to as "relevant other" (O) as opposed to the "self" (S). Specifying only one of the other players as O on any given trial, and changing this from trial to trial, also helped reduce sequential dependencies between performance estimates on successive trials. Whether or not a subject collaborates on any trial should depend on the subject's estimate of their own ability, their estimate of the other player's ability, and the threshold that has to be reached.

On other trials subjects were given the opportunity to compete the relevant other, O, in the next run of the minigame (Fig.3b). This meant that the difference between their own performance score and the other player's performance score would be taken and if it exceeded a threshold indicated on the screen then points were awarded. Subjects could first decide whether to engage in competition or whether to avoid it (engage/avoid decision). If they avoided competition then the points awarded were randomly distributed around zero. Again which of the two players was O varied from trial to trial. And again,

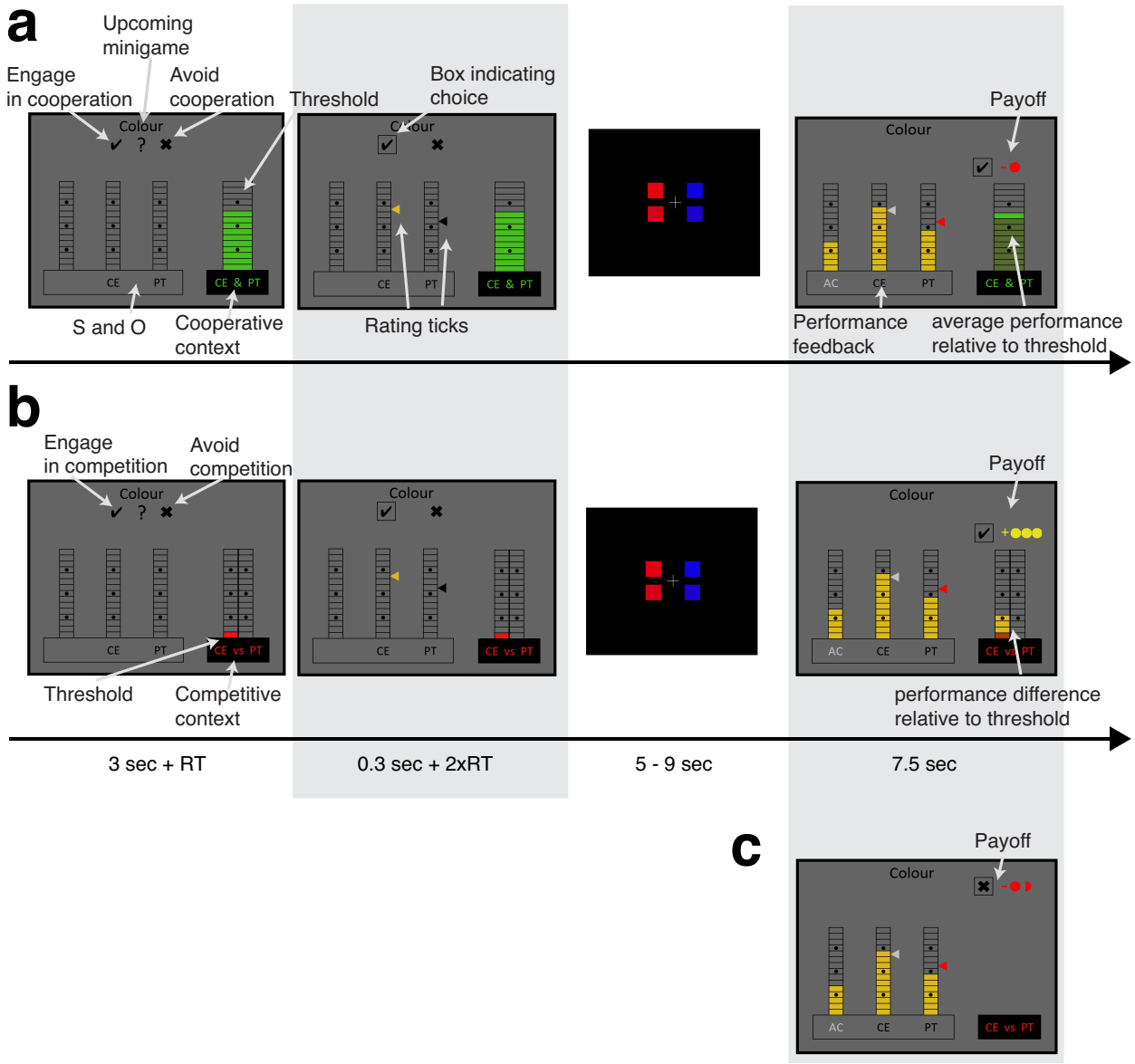


Figure 3. Trial time line. (a) *Trial timeline for an example cooperative trial. Trials start with the presentation of S, O, the upcoming minigame, social context and the threshold. A decision is made about whether to cooperate or not (in this case the subject decided to cooperate). After the decision, subjects provided, in randomized order, S-ability and O-ability ratings. Then all players play a minigame (independent of choice or ratings). In the feedback phase, the payoff (-1 in this instance) is determined by the disparity between the performance average of S and O (9) and the threshold they were asked to exceed (10). Payoffs were parametric and could be positive, zero, or negative.* ***(b)*** *Trial timeline for an example competitive trial. Annotations in the figure panel indicate elements differing from the cooperative example in panel a. Again, an engage choice is made but now the decision is whether or not to compete. In this case the subject chose to compete. Payoff (+3) is determined by the difference between the performance difference of S and O (4) and the threshold (1).* ***(c)*** *If subjects decided not to cooperate or to compete (avoid choices), they nonetheless went on to make ratings and perform minigames. In the feedback phase, the context was still displayed (competitive in this example), but no threshold was shown and the payoff was independent of performance scores. On such trials, subjects either won or lost 1.5 points with equal probability (i.e. the expected value was zero). Hence, choices should only depend on the expected value of cooperating or competing (depending on which context was indicated).*

whether or not a subject competes should depend on the subject's estimate of their own ability, their estimate of the other player's ability, and the threshold score that has to be reached.

In summary, the engage/avoid decision created either a cooperative or a competitive relationship between S and O; O became either an ally or an opponent for S on that trial (Fig.3). The context varied pseudorandomly across trials making sure that slowly drifting ability estimates were comparable between competitive and cooperative trials. Importantly, the inclusion of the engage/avoid decision and the threshold enabled us to distinguish S and O related brain activity from brain activity related to reward processing (see next chapter).

So that we could measure subjects' absolute ability estimates, subjects were asked on each trial to rate the performance they expected for the current trial for themselves (*S-ability*) and the relevant other (*O-ability*). *S-ability* and *O-ability* were assessed via two independent ratings (in random order) prior to performing the minigame. Having the ability ratings embedded in either a cooperative or a competitive context allowed us to test whether past performance of S and O had differential effects on the *O-ability* and *S-ability*, respectively. We used ratings in which subjects rated the players relative to a rating marker, which was updated over trials using a staircasing procedure (see "Ability ratings" in Methods) to maximize both informativeness and speed of the ratings (and therefore trial number and statistical power). The inclusion of two different other players who took turns to be the relevant other in a trial ensured that S's and O's performance feedback and associated ability estimates were statistically decorrelated allowing identification of neural correlates in an unconfounded manner (see next chapter).

4.2.3. *Detailed experimental design and schedule*

While lying in the magnetic resonance imaging (MRI) scanner, on each trial,

subjects performed a minigame. They were led to believe that two other subjects played the same minigame simultaneously. Each trial took place either in a cooperative or in a competitive social context. Subjects then made an engage/avoid decision. On some trials the choice was between cooperating or refraining from cooperating, while in other trials the choice was between competing or refraining from competing. If subjects took the "avoid" choice then that meant that they were simply awarded a small number of points (1.5 points) randomly distributed around zero (and they were informed that this was the case). However, if they took the "engage" decision in the cooperative context then they opted to ally themselves with one of the other players (which one varied from trial to trial and was indicated by the experimenter on each trial and is subsequently referred to as the *relevant other* - O) to see if together they could perform well enough for their average points to exceed a threshold level (which varied from trial to trial). If they did, they gained reward points on that trial, but if they fell short of the threshold they lost points. By contrast, if they took the "engage" decision in the competitive context then the other player became an opponent. The difference between the subject's and opponent's performances then had to exceed a threshold (again the threshold was variable) and the payoff was proportional to this difference (i.e. a win, a loss or neither of the two). In summary, the social context was critical when decisions to engage were made. Reward outcomes for engage/avoid choices were determined by minigame performances of S, O, and a threshold that varied unpredictably from trial to trial.

1a) $$EngagePayoff_{Competition} = (feedback_S - feedback_O) - threshold$$

1b) $$EngagePayoff_{Cooperation} = (feedback_S + feedback_O)/2 - threshold$$

"Feedback" in equation 1a and 1b refers to performance feedback. The timing of events within each trial is illustrated in Fig.1,3,6.

While the likely performance feedback for S and O could be estimated from

performance feedback on previous trials, the threshold varied unpredictably from trial to trial and was used to dissociate reward expectation from performance expectation and to make sure that subjects did not make their choices before the beginning of the current trial. Subjects found the meaning of the thresholds intuitive when the task was being explained to them and their task behavior confirmed that they had understood the task.

Subjects then also provided an estimate of their ability on each trial by rating the expected performance for themselves (S) and the *relevant other* (O) for the upcoming trial of the minigame. The order of S and O ratings was randomized across trials. As explained above, although both of the two other players performed the minigame simultaneously, subjects were only paired (to compete or cooperate) with one of the other players (the relevant other, O). Therefore only O, and not the third player, was relevant for a trial's engage/avoid decision. However, the identity of O switched between trials. On each trial, after the minigame, subjects received performance feedback about themselves as well as about the performances of the other two players. As minigames, we designed two short reaction-time based tasks. See Figure 4 and Fig.9 for details on the minigames. The goal of the subjects in the experiment was to collect as many rewards (points) as possible, as these were translated into monetary rewards at the end of the experiment.

For all three players, including the subjects themselves, performance feedback on every trial was predefined (Fig.4). In other words, the feedback about performance was independent from subjects' actual performance in the minigames (see, however, "false start trials" for a case of veridical performance feedback in the "Feedback" section below). This was necessary to control and match performance feedback between subjects as well as between subjects and the two other players. Subjects were told that the minigames had been tested on a larger sample of subjects and that performance feedback in the minigame reflected individual performance relative to that sample.

In the phases before and after the minigames, three scales ranging from 1-15 points

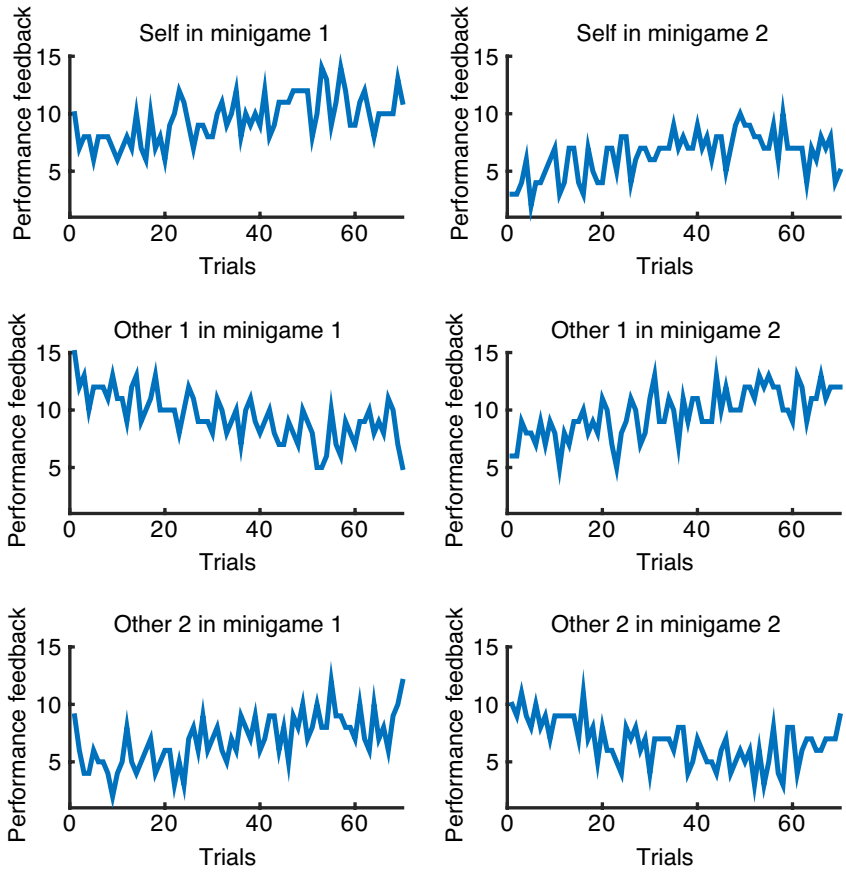


Figure 4. Performance feedback schedules. *Performance feedback was predetermined to guarantee that self and relevant other (S and O, respectively) performance feedback allowed a meaningful data analysis. The performance feedback for S and O was similar in mean and variance. Subjects performed one of two minigames on each trial and S performance feedback was more positive in one minigame compared to the other minigame (counterbalanced across subjects), while variance in S performance feedback was similarly small in both minigames. This was done to sample a wide range of S performance feedback while still keeping S performance feedback relatively stable within each minigame. The same logic applied to the performance feedback schedules of the two other players. In short, mean, variance and slow drifts of performance feedback were balanced for self and others to ensure that differential effects were not driven by schedule differences. The relevant other (O) in a trial was Other 1 in half of the trials and Other 2 in the other half. Trials with minigame 1 and minigame 2 were pseudorandomly interleaved.*

were shown with the initials of the three players below. Performance feedback was displayed on these scales in the feedback phase. While the initials of the confederates were the same for all experimental sessions, the subjects' own initials were adjusted to be appropriate for each individual subject. The initials created a social frame for the

experiment without using explicitly social cues such as faces.

The experimental schedule contained 136 minigame trials. The design was a 2[social context] x 2[partner] x 2[minigame] fully crossed design (17 trials per cell). This meant that a trial could be either cooperative or competitive [social context: cooperation or competition], the O could be either "player" 1 or 2 [O: Other1 or Other2] and, in each trial, subjects played one of two minigames [minigame: time task or color task]. The trial type order was pseudorandom and the same for all subjects. Three subjects performed a marginally shortened version of the schedule. They did the first 116 trials of the schedule, however the design was still fully crossed (14 trials per cell).

Several features of the experiment were counterbalanced independently across subjects to avoid confounds. The mapping between minigame (time task or color task) and associated performance feedback schedule was counterbalanced. Moreover, as the left/right side of the buttons used to indicate the engage/avoid choice were fixed for each subject, they were counterbalanced across subjects. We also counterbalanced the screen location where initials of each confederate were displayed (left or right of subjects' initials which were always in the middle), and association between confederates (whom the subjects had met before going into the scanner) and their performance feedback schedules.

Thus, in summary, all trials comprised an engage/avoid decision, two binary ratings (for S and O), a minigame phase and a feedback phase. Timing details for all phases (except minigame phases) are shown in Fig.6.

4.2.4. *Ability ratings*

As already mentioned subjects provided S and O ability ratings. For each rating, initially, a tick indicated a value on the performance scale (rating marker) and subjects indicated if expected performance (for S or O as appropriate) would surpass or fall below the rating marker (Fig.5). A positive rating (i.e. performance is expected to be above the

rating marker) was made with one button and a negative rating (i.e. performance is expected to be below the rating marker) was made with the other button. As performance feedback was always expressed in integers, the rating makers were always set between two integers (X.5-values). The rating marker was updated from trial to trial based on the rating choices for the respective player using a staircasing procedure to increase sensitivity of the ratings. A positive rating resulted in an increase of the rating marker's value by one point in the next trial of the same minigame for the given player; a negative rating resulted in a decrease by one point. The value of the rating marker on the first trial after the starter trials (see below) was based on the player's mean performance feedback in the starter trials.

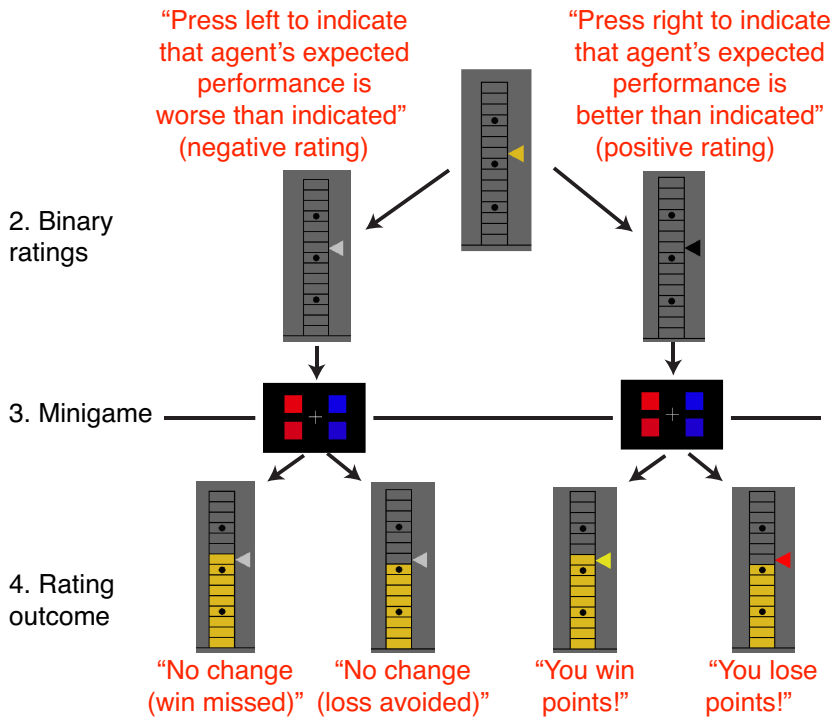


Figure 5. Rationale of ability ratings. *Before the minigame, subjects indicated for S and O either a positive or a negative rating reflecting the expectation that the player would surpass or fall below a given rating marker. The color change of the rating marker was indicative of the choice made. The rating marker turned black for a positive rating and grey for a negative rating. A positive rating led to a win or loss of 0.25 points depending on subsequent performance feedback. A negative rating led to no change in the points count independent of performance feedback. Therefore, making a correct negative rating was associated with a benefit of avoiding losing points while making a correct positive rating was associated with a benefit of winning 0.25 points. Red text in quotes is taken from the subjects' instructions where a similar illustration was used.*

Subjects received a small payoff for the rating that depended on the accuracy of the ratings they gave prior to playing a minigame. However, to reduce any incentive to perform badly on the minigames, negative ratings never yielded payoffs. For positive ratings, however, subjects won or lost 0.25 points depending on whether the subsequent performance feedback received surpassed or fell below the rating marker. Note that the magnitude of the rating payoff was insignificant compared to the payoff for the engage/avoid decision.

4.2.5. *Feedback*

Feedback at the end of each trial was chunked together in three components and the components were presented in a randomized order. The first component was the performance feedback for S and O, which were presented simultaneously with the information about the accuracy of the rating subjects had made (change of rating marker color; see Fig.5). The second component was the payoff of the engage/avoid decision. For this, a cue indicating the trial's choice appeared on the right side of the screen (a tick for engage choices and a cross for avoid choices) together with circles representing coins that were won (yellow circles with a plus sign) or lost (red circles with a minus sign). At the same time, only for engage choices, the performance feedback average (cooperation trials) or performance feedback difference (competition trials) were displayed on the scales on the right side of the screen. The third component was the performance feedback of the other player that was not the O (irrelevant other). The initials of this player were displayed in a different color and the performance was irrelevant for any payoff. The three feedback components appeared in random order to control for sequence effects and stayed on screen once they were presented. The first component occurred after one second, the second component 1.25 seconds later, the third component another 1.25 seconds later and the feedback phase ended after a further four seconds. Then, after 0.5-2.5 second inter-trial-

interval (ITI) with a blank screen, the next trial started.

There were two types of trials which deviated from the described structure. First, the first four trials of the experiment were "starter trials" (two with the time task, two with the in color task). Those trials were for subjects to form initial ability estimates about themselves and the other players. For this reason, in starter trials, there was no option to cooperate or compete) and no ability ratings were made. Second, for trials where subjects performed very badly in a minigame ("false starts") the feedback phase was adjusted. The performance thresholds for false start trials are discussed below. In false start trials, subjects received no performance feedback for themselves, but only for the other players (Fig.14b in the results section shows the number of false start trials per subject). The sole payoff for false start trials was a loss of two points independent of subjects' ratings and engage/avoid choice. Subjects were instructed about this. It was explained that extremely bad performances would be detected by the computer running the experiment and discarded as false starts to sort out performance slips that were, for instance, due to inattentiveness and did not reflect a player's "true" performance. This procedure was used to make the pre-determined feedback in other trials more believable as the feedback in false start trials was actually determined by true minigame performance. Note that subjects were also told during the instructions that there would be a special type of false start trial if one of the other players performed very badly. However, this never happened in the experiment. Starter trials and the feedback phase of false start trials were excluded from fMRI analysis.

A detailed time trial time line including all trial elements and timings is shown in Figure 6.

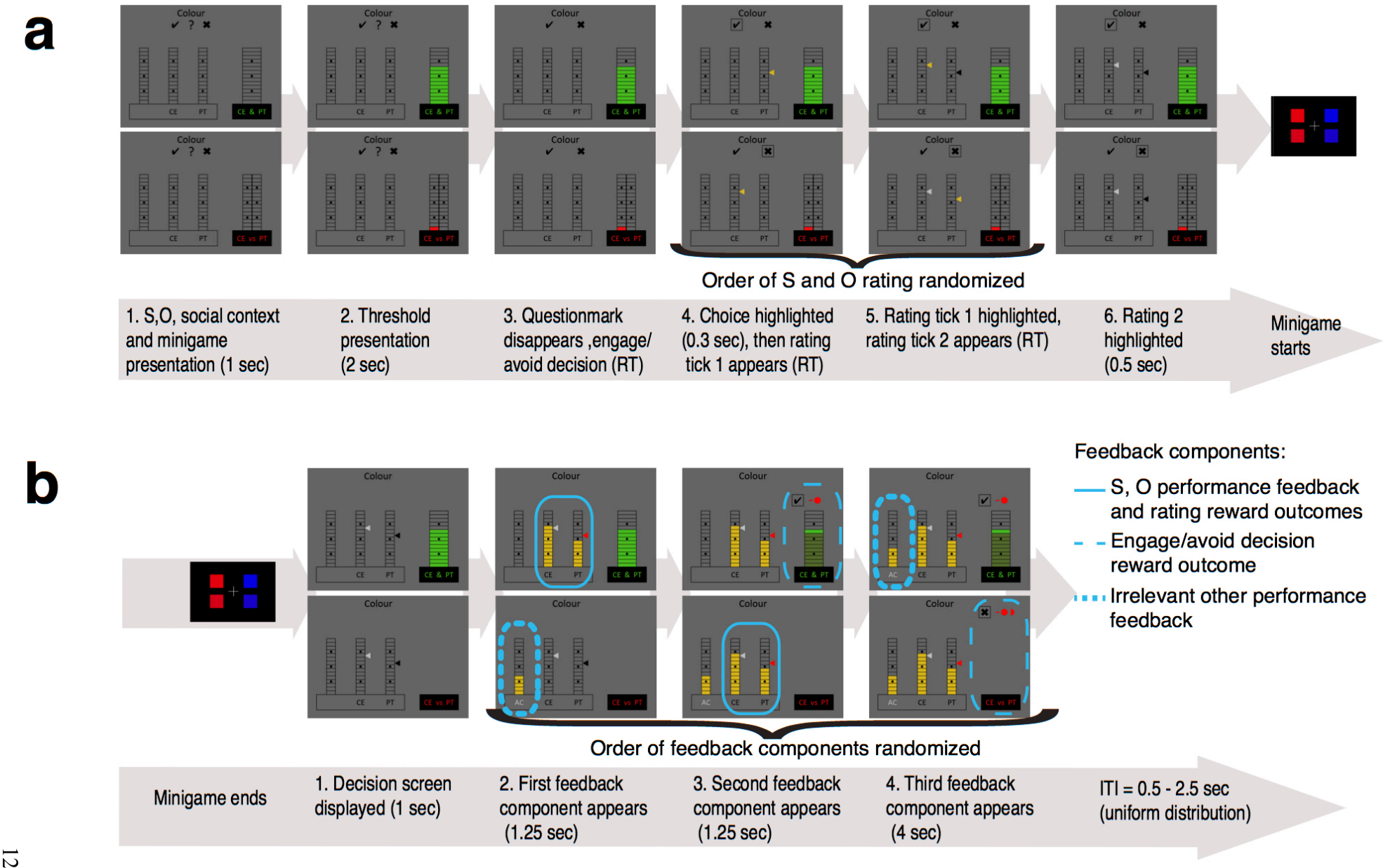

a
Order of S and O rating randomized
1. S,O, social context and minigame presentation (1 sec)
2. Threshold presentation (2 sec)
3. Questionmark disappears ,engage/ avoid decision (RT)
4. Choice highlighted (0.3 sec), then rating tick 1 appears (RT)
5. Rating tick 1 highlighted, rating tick 2 appears (RT)
6. Rating 2 highlighted (0.5 sec)
Minigame starts
b
Feedback components:
S, O performance feedback and rating reward outcomes
Engage/avoid decision reward outcome
Irrelevant other performance feedback
Order of feedback components randomized
Minigame ends
1. Decision screen displayed (1 sec)
2. First feedback component appears (1.25 sec)
3. Second feedback component appears (1.25 sec)
4. Third feedback component appears (4 sec)
ITI = 0.5 - 2.5 sec (uniform distribution)

Figure 6. Detailed trial structure (see last page). *For an example trial, we show the first part of the trial in which the social context (competition or cooperation) was revealed, participants made the engage/avoid decision, and provided self (S) and relevant other (O) ratings (a). The feedback phase is shown of the same example trial (in b). For illustration purposes, the same performance feedback is shown in a cooperative context (upper rows) and a competitive context (lower rows). Note that timing of the events are shown in brackets and RT (reaction time) means that the respective step only ends when the subject makes a response.* ***(a)*** *At the beginning of a trial S (initials of the subject), O (initials of one of the two other players relevant for the current trial), social context (cooperation or competition) and type of upcoming minigame ("Colour" or "Time") is presented. The initials of the irrelevant other player are not shown and the social context is color-coded indicating also the threshold of the current trial. In these examples, the choices made (see highlighted tick and cross in step 4) are to engage in cooperation (indicating the expectation that S and O will, on average, perform better than the threshold of 10) and to avoid competition (indicating that the subject does not expect to perform at least one point better than the O). After the engage/avoid choices, binary ratings of S and O take place (step 4 to 6) in randomized order (in these examples, O rating first in cooperation and S rating first in competition). The ratings shown here indicate that the subject expects himself or herself to perform worse than 10.5, while the O is expected to perform better than 8.5 (see Fig.5 for color coding of rating ticks). Note that these ratings are consistent with the engage/avoid choices shown in the two examples.* ***(b)*** *In the feedback phase, the previous screen from before the minigame reappears. However, the right side of the screen showing the cooperation/competition threshold is occluded if a subject had chosen to avoid cooperation/competition in the previous decision phase of the trial. This means that in this example trial the threshold is only shown again in the cooperate trial (because the subject actually decided to cooperate) and not in the compete trial (because in this example trial the subject refrained from competing). In other words, the repeated presentation of the threshold is not a feature of the social context, but entirely a consequence of the engage/avoid choice made. Subsequently, three feedback components appear in randomized order to control for sequence effects (see legend on the right hand side). In cooperation, the choice payoff is -1, because the average performance is 9 while the threshold is 10. In competition, the choice payoff is -1.5 which is due to chance (payoff from avoid choices is +1.5 or -1.5 with a 50/50 probability) and independent of performance feedback. Note that the subject would have earned a payoff of 3 had the engage choice been taken (performance feedback difference of +4 minus threshold of +1). Note that, overall, the magnitude of the rating payoff is marginal compared to the engage/avoid choice payoff.*

4.2.6. *Experimental procedure, instructions and debriefing questionnaire*

The experimental procedure was precisely scripted and involved two experimenters, two confederates, and the radiographer to make subjects believe they would be playing an interactive game together with two other subjects. The same two confederates pretended to be the two other players for every subject. They arrived in the laboratory at a similar time as the subjects. Subjects had been told in advance that the

experiment investigates learning and decision making in social situations and that two other naïve subjects would participate in the same experiment. After a short introduction, subjects were separated from the confederates and instructed about the experimental task. They were told that they would be playing an interactive game together with the confederates, who would play the game from computers outside the scanner room. A faked lottery was used to determine who of the three subjects went in the scanner. For approximately 45 minutes, subjects were instructed on the minigames, the ratings and the engage/avoid decisions and performed example trials.

The instructions were designed such that actual performance learning would only take place in the subsequent experiment to maximize learning effects in the scanner. To still guarantee familiarity with the minigames, a written explanation of the two minigames was complemented with a short practice session, in which subjects performed 5 trials of each minigame (Fig.7a). During these trials, the experimenter was present and made sure that subjects understood the minigames. No explicit performance feedback was given on those trials to avoid performance learning. To guarantee that subjects understood the logic of the engage/avoid decision and the ratings, subjects performed 16 example trials that did not include minigames, but instead a placeholder screen (Fig.7b). This allowed subjects to adjust to the trial events and experience the reward outcomes of ratings and decisions. Importantly, the performance feedback on those trials followed no across-trial contingencies and consisted mostly of the highest or lowest performance feedback for the players. This was done to make these example trials very different from the trials experienced in the main experiment. Although subjects could not learn anything during those trials and therefore could make no well-grounded decisions, they were asked to invent and verbalize reasons for their ratings and decisions so that the experimenter could make sure that they understood their logic (e.g. "This is a cooperate trial. I press the "engage" button, because although the threshold is high, I think we will perform very well.

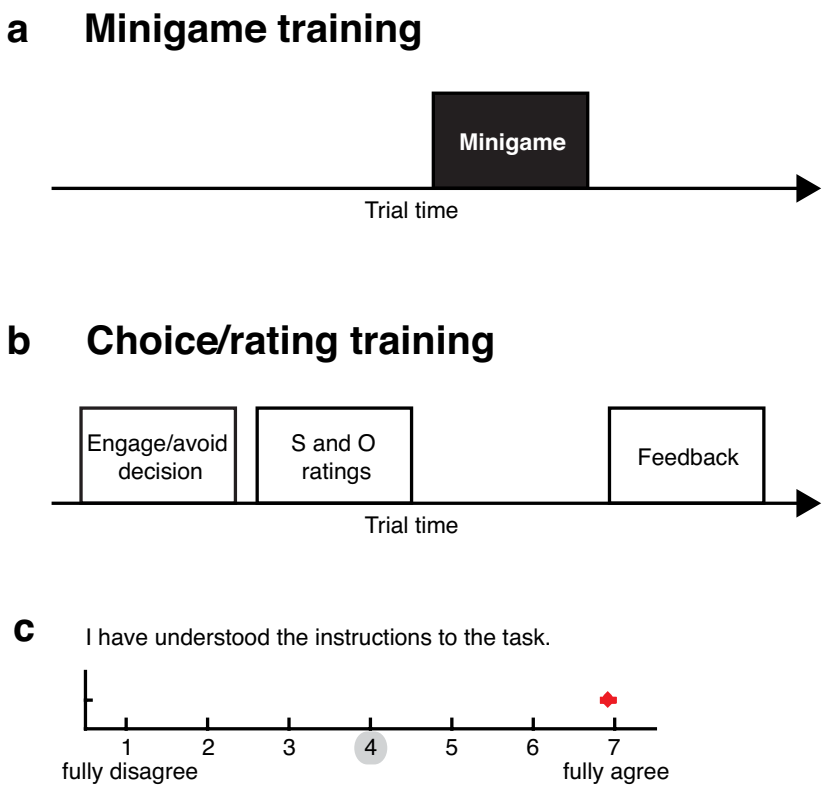


Figure 7. Instructions. *(**a**) Subjects first experienced some trials of the minigames without choice feedback. (**b**) Afterwards, subjects performed some trials of the experiment where the minigames were not played to familiarize them with the ratings and choices. (**c**) In our debriefing questionnaire after the experiment, subjects confirmed that they had understood the task instructions. 22 subjects gave a 7 out of 7 rating for the question "I have understood the instructions to the task" (the remaining 2 subjects gave 6 out of 7). (small horizontal error bar is +/- SEM).*

I rate myself and the other one positively, because I think we will both perform well...").

In sum, subjects practiced all aspects of the experimental task, but in such a manner that they could not yet learn about their performance. Subjects were told that their goal in the experiment was to collect as many points as possible and that points would be translated into monetary reward at the end of the experiment. It was emphasized that points could be earned by making good decisions and by providing accurate ratings of performance. Despite the substantial time needed for a thorough instruction, most subjects found the task intuitive and both the behavioral data acquired in the experiment and a post-experiment questionnaire confirmed that they understood the task (Fig.7c).

The subsequent experiment in the MRI scanner took approximately 55 minutes.

After functional and structural MRI sessions, subjects filled in two short questionnaires and were then fully debriefed about the experiment.

No subject indicated any suspicion about false performance feedback or the identity of the confederates before debriefing. The results of the debriefing questionnaire supported this (Fig.8). The scales of the questionnaire were anchored for highest and lowest scores (see Fig.8). These data showed that across the whole sample, subjects did not rate their overall performance in the two minigames differently (item 1; paired t-test: t_{23}=0.13; p=0.9;) and that neither minigame performance feedback was rated as matching their subjective performance better (item 2; t_{23}=0.78; p=0.44). This was expected as we counterbalanced across subjects whether time task or color task was used as minigame 1 which had higher performance feedback on average compared to minigame 2 (Fig.4; for 11 of 24 subjects the time task was used as minigame 1). Note also that there was no confound in the ordering of minigames because trials of minigame 1 and minigame 2 were interleaved across the whole experiment. When we retrospectively recoded item 1 and item 2 with respect to minigame 1 and minigame 2, we found that subjects rated themselves better in the minigame where they received better performance feedback (item 1*; t_{23}=7.72; p=7.7*10^{-8}) and that the performance feedback for this minigame matched better how they subjectively felt about their performance (item 2*; t_{23}=4.03; p=5.2*10^{-4}). Note that the performance scores for item 1 and 1* are very similar to the actual average performance feedback (compare with Fig.4) indicating further that subjects learned from the performance feedback. To confirm the believability of the performance feedback, we asked subjects whether they felt that the performance feedback was accurate. If subjects experienced a mismatch between their subjective experience and the performance feedback they received, this would have been indicated by “disagree” ratings. We found slight agreement, however, indicating that subjects were unable to tell the inaccuracy of the performance feedback from their subjective experience. Additional suspicion checks asked

whether they tried to perform better than O in competition (indicating that they believed that they could gain positive performance feedback by trying hard; item 3) and if positive performance feedback was valuable for them (item 4). In all cases (items 2,3,4) subjects tended to agree rather than disagree (average scores higher than scale point 4). This suggests that subjects indeed experienced the performance feedback as a consequence of their actions in the minigames. Note, importantly, that in addition to the debriefing questionnaire, a debriefing interview took place for every subject with additional suspicion

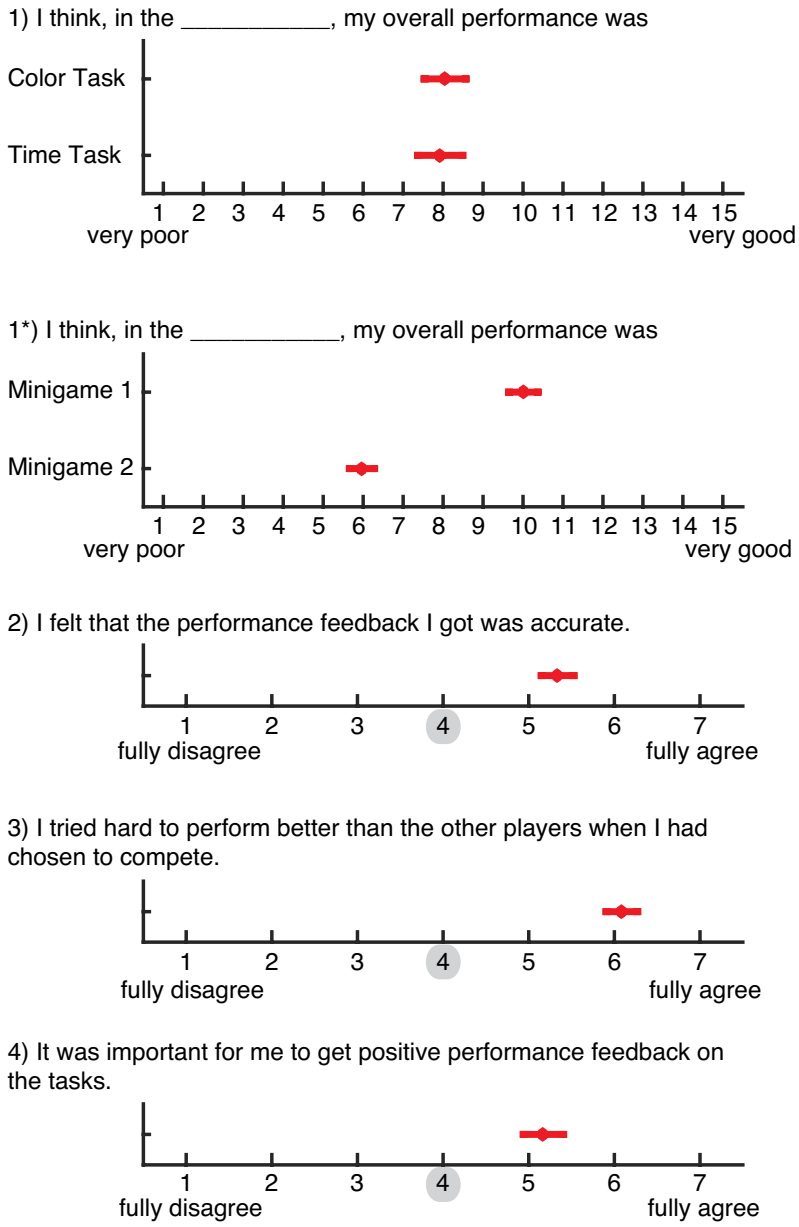


Figure 8. Debriefing questionnaire. *(error bars are +/-SEM across subjects).*

checks. As we handed out this questionnaire before debriefing, to avoid biased responses, we could only indirectly ask if subjects had suspicions about the deception used. Therefore, we conducted debriefing interviews to fully disclose the purpose and method of the experiment to subjects, but also to confirm that all subjects, without exception, indicated that they believed the false performance feedback and had no doubts about the identity of the confederates.

Note that, overall, we have several indications (and no counter-indications) that subjects' ability estimates were guided by the performance feedback and therefore that subjects found the feedback credible: the effects of past performance feedback on i) *S-ability* and *O-ability* ratings, ii) on decisions to engage in cooperation/competition (both in results section below, Fig.10,11), iii) on true minigame performance (see results section Fig.13); and iv) subjects' self-reports of the feedback credibility in a debriefing interview, documented by a debriefing questionnaire (Fig.8).

4.2.7. *Minigames*

We used two minigames in the experiment labeled "time task" and "colour task". The minigames were necessary prerequisites to administer performance feedback on every trial. They were short reaction-time based tasks and we designed them to be relatively non-transparent in the sense that the subjective experience of performing the minigames was not very informative for estimating one's ability compared to the explicit performance feedback that was given to the subjects. Also, we varied the timing parameters of the minigames to make it harder to compare true performance across different trials of a minigame. Lastly, we designed both minigames so that they appeared to be measuring different skills to make it plausible for subjects that performance levels in one task were not predictive of performance level in the other task. Each minigame was performed with one response button press with the right index finger.

The color task was (Fig.9a) loosely based on perceptual decision making tasks (Michael et al., 2014). Two pairs of squares, one pair red on the left, one pair blue on the right of a fixation cross, initially appear on screen. Then, the squares gradually change color over several seconds until the pairs have reversed their color. The color change occurs at an uneven rate. Subjects were asked to press a button when both pairs of squares had the same average color. The average colors were defined in RGB space as:

$$Color_{left} = [c_{red} \quad 0 \quad 1 - c_{red}]$$

$$Color_{right} = [c_{blue} \quad 0 \quad 1 - c_{blue}]$$

The colors were controlled by two parameters c_{red} and c_{blue}. For the left, initially red ([1 0 0] in RGB space) side, the parameter c_{red} decreased linearly from 1 to 0 over a run of the minigame. Vice versa, for the right, initially blue side ([0 0 1] in RGB space), the parameter c_{blue} increased linearly from 0 to 1 over a run of the minigame. The optimal time to press was when c_{red} and c_{blue} had the same value (color equivalence point), which happened only once as both parameters changed in opposing directions. The time of color equivalence was set to occur at a randomly picked time point drawn from a uniform distribution between 2.5 seconds and 4.7 seconds after onset of the color change. The c-values at that time were set to have a random value ranging from 0.25 to 0.75, also drawn from a uniform distribution. This meant the color equivalence point could lie rather towards the pure red or the pure blue side (a value of 0.5 indicates an even mix). Note that c_{red} and c_{blue} characterize the average color of the two left and two right squares, respectively. Hence, c_{red} (and c_{blue} analogously) was subdivided in c_{red1} and c_{red2} by multiplying c_{red} with a scaling parameter "colscale", which was randomly picked for each run of the minigame and ranged between 0.5 and 1.5 ($c_{red1} = c_{red}$ x colscale; $c_{red2} = c_{red}$ x (2-colscale)). In other words, c_{red} was subdivided so that it was always the average color of c_{red1} and c_{red2}. Performance on this task was calculated as:

$$Performance_{ColorTask} = (c_{blue} - c_{red}) \times 100$$

Optimal performance was represented by a performance value of 0 indicating a color match and the performance values represented deviation from color equivalence in percent. Negative performance values indicate a response that was too quick (the point of equal c values had not yet been reached when the subject responded) and positive performance values indicated a response that was too slow (c values are past the point of equality). If absolute performance values exceeded 60, trials were classified as false starts. This threshold was based on pilot experiments. Note again, that except for false start trials, the performance feedback was unrelated to the performance measure explained here.

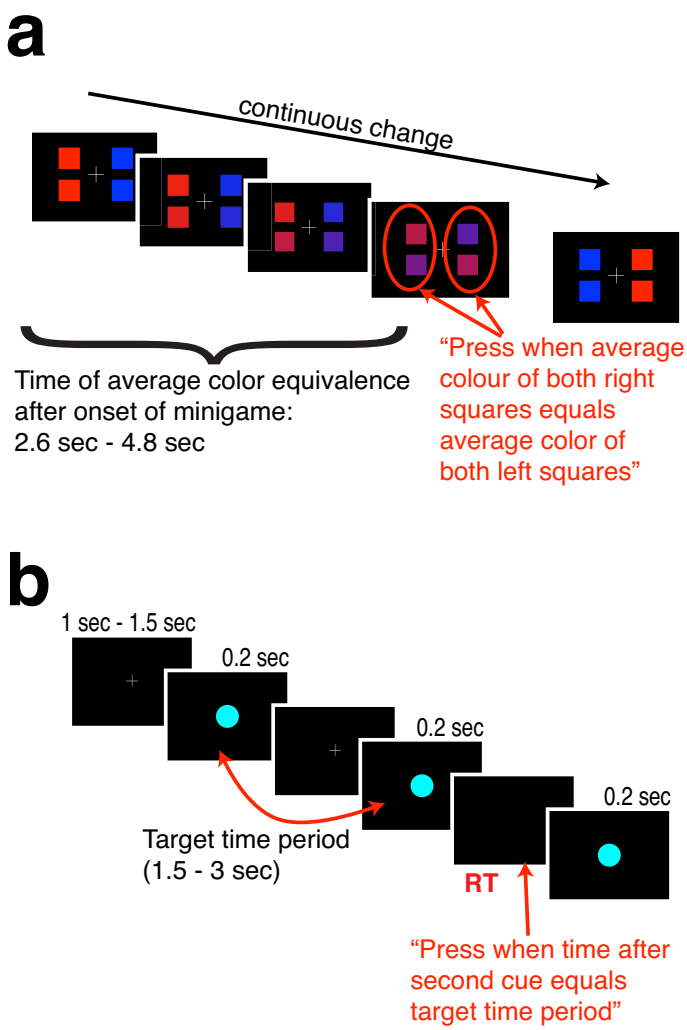


Figure 9. Minigames. *On half of the trials, subjects performed the "color" minigame (**a**) and on the other half the "time" minigame (**b**). Red text in both panels is taken from the subjects' instructions where a similar illustration was used. (RT, reaction time).*

In the time task subjects had to replicate a given time gap between two short occurrences of a blue dot on the screen (target time; Fig.9b). The target time was randomly picked for each trial from a uniform distribution and ranged between 1.5 and 3 seconds. Subjects were instructed to press when the time elapsed after the second dot appearance was equal to the target time. The response time was the time from the second dot appearance to the button press. Therefore, performance on the task was calculated as:

$$Performance_{TimeTask} = (response\ time / target\ time) \times 100 - 100$$

As in the color task, optimal performance on the time task was represented by a performance value of 0 and the performance values represented deviation from the optimal response time in percent. Positive performance values indicated a response that was too slow and negative performance values indicated a response that was too quick. If absolute performance values exceeded 70, trials were classified as false starts. This threshold was based on pilot experiments.

4.2.8. *Reinforcement learning model*

For every subject, we fitted a standard reinforcement learning model to model the performance estimates assigned to the three players trial by trial (Self, S; Other1, O1; Other2, O2). We used two minigame-specific (Color Task, T1; Time Task, T2) performance estimates per player. The performance estimates summarize the previous performance history and are hence referred to as *performance*. They reflect the expected performance based on a recency-weighted average of past performance feedback. This resulted in six player and minigame specific *performance* estimates: $performance_{S\text{-}T1}$, $performance_{S\text{-}T2}$, $performance_{O1\text{-}T1}$, $performance_{O1\text{-}T2}$, $performance_{O2\text{-}T1}$, $performance_{O2\text{-}T2}$. On every trial t, the three *performance* estimates associated with the current minigame were updated using a prediction error (PE) based learning rule with a learning rate α as a free parameter:

1) $Performance_{t+1} = performance_t + \alpha \times PE_t$

(formula was applied separately for S, O1, O2, given T1 or T2)

The PE itself was calculated based on the specific *performance* estimate and performance feedback of the player in the current minigame as:

2) $PE_t = feedback_t - performance_t$

(formula was applied separately for S, O1, O2, given T1 or T2)

In false start trials, the *performance* estimate for S was not updated and remained unchanged until the next trial of the same minigame. No PEs for S were calculated for false start trials (the other players never displayed false start trials). For the first trial of the session for each player for each minigame, no *performance* estimates and prediction errors were calculated, but the initial performance feedbacks were used as the starting *performance* estimates for the respective player in the respective minigame.

In each trial after the four starter trials, subjects made a decision about cooperating or competing (depending on the current context) with the relevant other (O) and in addition provided ratings of both S and O. Both engage/avoid decisions and ratings were modeled based on *performance* estimates for S and the O, called *S-performance* and *O-performance*. *S-performance* is the *performance* estimate for Self associated with the minigame of the current trial ($performance_{S\text{-}T1}$ or $performance_{S\text{-}T2}$). Similarly, *O-performance* refers to $performance_{O\text{-}T1}$, $performance_{O1\text{-}T2}$, $performance_{O2\text{-}T1}$ and $performance_{O2\text{-}T2}$, depending on which other player was currently selected as the O and which minigame took place. Therefore, *S-performance* and *O-performance* represented minigame- and player-specific performance expectations of the players involved in the current trial's engage/avoid decision. The same was the case for the PEs associated with S and O.

Subjects' ratings of a player reflected expectations of whether they would perform either better or worse than a level indicated by a *rating marker* the position of which was adjusted from trial to trial using a staircase procedure (see "Ability ratings"). Expectations expressed in the ratings that exceeded or fell below the rating marker were referred to as positive and negative ratings, respectively. To calculate the probability of a positive rating (p(positiveRating)), we used a softmax function with an inverse temperature β. This was done separately for S and O using *S-performance* and *O-performance*, respectively as well as the player-specific rating marker:

$$3)\ P(positiveRating) = \frac{\exp[\beta \times (performance - ratingmarker)]}{\exp[\beta \times (performance - ratingmarker)] + 1}$$

(this formula was applied separately for *S-performance* and *O-performance* given their respective rating markers)

Having calculated the probability of a positive rating on a given trial, the probability of the rating actually observed was derived, again, separately for S and O:

$$4)\ P(rating) = \begin{cases} p(positiveRating) & if\ rated\ positively \\ 1 - p(positiveRating) & if\ rated\ negatively \end{cases}$$

(formula was applied separately for S and O)

Subjects also received a small gain or loss at the end of a trial if they had made a positive rating and the expectation indicated by that rating had been accurate (ratingbonus of 0.25 points). As explained in the Methods, to ensure that there was no temptation to perform poorly in the task no ratingbonus was awarded when a negative rating had been given. The expected value of a rating (EV_{rating}) was calculated as

$$5)\ EV_{rating} = \begin{cases} [p(positiveRating) - 0.5] \times 2 \times ratingbonus & if\ rated\ positively \\ 0 \quad if\ rated\ negatively \end{cases}$$

(formula was applied separately for S and O)

Note the bounds of EV_{rating} for positive ratings are 0.25 and -0.25, which are the points that can be lost or won for positive ratings.

In addition to completing a rating for S and O on each trial, subjects made a decision to engage in or avoid cooperating/competing. Given the objective social-context-specific payoff scheme of the task (equations 1a and 1b in "Detailed experimental design and schedule" above), the subjective expected value of engaging in cooperation/competition (EV_{engage}) was calculated in an analogous way:

6a) $Competition: \quad EV_{engage} = S\text{-}performance - O\text{-}performance - threshold$

6b) $Cooperation: \quad EV_{engage} = (S\text{-}performance + O\text{-}performance)/2 - threshold$

A decision to avoid cooperating/competing led to a gain of 1.5 points and a loss of 1.5 points with equal probability and subjects had been instructed that the expected value of the decisions to avoid cooperating/competing was zero:

7) $$EV_{EAD} = \begin{cases} EV_{engage} & if\ engage \\ 0 & if\ avoid \end{cases}$$

Therefore, EV_{engage} was used as decision variable for the engage/avoid decisions to calculate the probability of engaging in cooperation or competition:

8) $$P(engage) = \frac{\exp(\beta \times EV_{engage})}{\exp(\beta \times EV_{engage}) + \exp(\beta \times EV_{avoid})}$$

Note that EV_{avoid} is zero in equation 8, as explained above. The probabilities of the actual choices made were derived from p(engage):

9) $$P(choice) = \begin{cases} P(engage) & if\ engage \\ 1 - P(engage) & if\ avoid \end{cases}$$

The full reward expectation on each trial (EV_{chosen}) was defined as the sum of the expected values from both ratings and the expected value of the engage/avoid decision (equations 5 and 7):

10) $EV_{chosen} = EV_{S-Rating} + EV_{O-Rating} + EV_{EAD}$

The reward prediction error (RPE) was calculated based on all reward outcomes of a trial including both rating reward outcomes and the engage/avoid decision reward outcome (see above equation 2 for the calculation of player specific prediction errors):

11) $RPE = \text{Reward} - EV_{chosen}$

Overall, the free parameter set θ comprised two free parameters: the learning rate α and the inverse temperature β. We fitted these parameters for every subject separately by minimizing the negative log likelihood (nLL) over all trials N, given a set of parameter values. For the calculation of nLL, we treated ratings and engage/avoid decisions equally. So the decisions used to fit the model included equal proportions of ratings of S, ratings of the O and engage/avoid choices to improve the model fit.

12) $nLL = -\sum_{n=1}^{N} \log(p(decision_t | \theta))$

4.2.9. *Behavioral analyses*

We used general linear models (GLM) to examine the impact of different factors on behavior. GLMs were applied to the ability ratings for S and O ("Rating GLM for S/O"), to the engage/avoid choices ("Choice GLM") and to an index of true performance change over trials ("pChange GLM"). For all GLMs, all regressors were normalized (mean of 0 and standard deviation of 1).

Our first analyses aimed to show that subjects relied on performance feedback for their ratings and engage/avoid choices. For this, we predicted S and O ratings on the basis of S's and O's recent performance feedback history (last four trials; referred to as "feedback" below) at trial t. In addition, the GLMs contained the value indicated by the rating marker (the rating marker indicated a performance level with reference to which the rating should be made; see "Ability ratings"):

Rating GLM 1 for S:

feedback-S_{t-1},feedback-S_{t-2}, feedback-S_{t-3}, feedback-S_{t-4}, ratingmarker-S (Fig.10b-i)

Rating GLM 1 for O:

feedback-O_{t-1}, feedback-O_{t-2}, feedback-O_{t-3}, feedback-O_{t-4}, ratingmarker-O (Fig.10c-i)

As an aggregate index of performance feedback, we fitted a reinforcement learning (RL) model on the rating data and used *S-performance* (recency-weighted performance estimate for self) and *O-performance* (recency-weighted performance estimate for relevant other) to explain engage/avoid decisions. For this analysis, the RL model was fitted on the ratings only and not on the engage/avoid decisions (all other analyses were fitted on both ratings and choices to improve model fit). We applied the same GLM separately to cooperative and competitive trials:

Choice GLM1 - binned by social context (cooperate/compete):

S-performance, O-performance, threshold (Fig.11)

The second set of analyses aimed to investigate more subtle effects of self-other-mergence (SOM) in the rating data. Again, we used *S-performance* and *O-performance* from the RL model. Note that the SOM GLM effects are orthogonal to the fitting done in the RL model and all SOM related effects presented can be shown without the use of a reinforcement learning model i.e. by using past performance feedback directly to explain choices (Fig.12b,13b,c). The following analysis was restricted to engage trials only (a supplementary analysis with the same results using all trials including those on which subjects refrained from cooperation and competition is shown in Fig.12c). We focused first on engage trials because the social context is critical on these trials (rather than on "avoid" trials when subjects simply took the default option of a random payment). The GLM was applied separately to cooperate and compete trials:

Rating GLM2a for S - binned by social context (cooperate/compete):

S-performance, O-performance, ratingmarker-S (Fig.10b-ii)

Rating GLM2a for O - binned by social context (cooperate/compete):

O-performance, S-performance, ratingmarker-O (Fig.10c-ii)

Note we used *S-performance* in the S rating and *O-performance* in the O rating as control parameters and the resulting beta weights for those regressors were only used as indices of individual variability, for instance to establish correlations with brain signals (see next chapter); we did not test such effects for significance as they had been fitted on the same ratings. The main aim of this GLM was to assess whether ratings indicated, in addition to agent-consistent effects, an inappropriate influence of the performance history of the different player (underlined above; for example an influence of *O-performance* on the S rating, or an influence of *S-performance* on the O rating). The binned rating GLM2a was used for visualization. However, as we were interested in the differences of agent-misattributed effects in cooperate and compete conditions (i.e. the interaction by social context), we devised an analogous rating GLM2b that was not binned by social context, but instead contained interaction terms of regressors of interest with social context. These interaction effects were tested for significance of SOM_{int} ("int" denotes "interaction"). To calculate the interaction terms, *S-performance x Context* for example, *S-performance* was normalized and multiplied with 1 for cooperate trials and -1 for compete trials (the number of cooperate and compete trials was identical over the whole session); the same was done for *O-performance*. Hence, a positive interaction effect indicated that the effect of S-performance was stronger in cooperation than in competition. For analyses restricted to engage trials the -1/1 term indicating cooperation or competition was normalized as well to account for differences in trial number. GLM2b, like GLM2a, was restricted to engage trials:

Rating GLM2b for S - with interaction by social context:

S-performance, O-performance, S-performance x Context, O-performance x Context, Context, ratingmarker-S (SOM_{int}(O-->S) is underlined; Fig.12a)

Rating GLM2b for O - with interaction by social context:

S-performance, O-performance, S-performance x Context, O-performance x Context, Context, ratingmarker-O (SOM_{int}(S-->O) is underlined; Fig.12a)

Additional versions of rating GLM2b, which do not use an RL model (Fig.12b) or are based on all trials (instead of engage trials only; Fig.12c) are shown in Fig.12.

We also analyzed performance in the minigames. For these analyses, we used a measure of subjects' true performance change (S-pChange) from any trial, t, to the following trial of the same minigame, t+1. S-pChange for a given trial t was calculated as the logarithm of the absolute performance difference:

$$S\text{-}pChange_t = \log(1 + |performance_{t+1} - performance_t|)$$

The performance measure was signed with zero indicating optimal performance while positive and negative values indicated responses that were too slow or too quick, respectively (Fig.14). The use of predetermined performance feedback schedules meant that performance feedback could not be improved by more optimal performance in the minigames (except false start trials as a case of veridical performance feedback, see Fig.14b showing the number of false start trials per subject). This made it possible to use S-pChange as a measure of behavioral adjustments based on performance feedback. Performances from false start trials were treated as outliers and therefore no S-pChange for a false start trial and the trial directly preceding it were calculated. Note that the performance measures for both minigames were comparable and the performance feedback schedules were balanced over both minigames. The main aim of this set of analyses was to

investigate whether behavioral adjustments were also influenced by performance feedback for the O depending on the social context (cooperation or competition). For this, similar to rating GLM2, we used two analogous versions of a GLM; one applied to cooperate and compete trials separately for visualization, and one version applied to all trials to test interaction by social context effects for significance. We investigated the relationship of S-pChange (for trial t indicating the transition from trial t to t t+1, see definition of S-pChange above) to the three variables that determine reward outcomes in engage/avoid decisions: S performance feedback, O performance feedback and threshold (all for trial t)

S-pChange GLM1a - binned by social context on trial t (cooperate/compete):

S-feedback, O-feedback, threshold (Fig.13b)

As in previous analyses, significance of effects was calculated over all trials, using the interaction effect of performance feedback and social context on trial t.

pChange GLM1b - all trials:

S-feedback, O-feedback, S-feedback x Context, O-feedback x Context, Context, threshold (SOM-related effect is underlined; Fig.13c)

Note that when using the threshold regressor over all trials, we combined the threshold regressors from cooperate and compete conditions and normalized them separately for each condition, as in the following neural analyses (see next chapter). The resulting regressor was also again normalized to have a mean of zero and standard deviation of 1. S-pChange was normalized separately for individual subjects in these GLMs. Positive interactions by context again indicated stronger influence of the performance feedback on cooperate compared to compete trials.

4.3. Results

4.3.1. *Appropriate ability estimation and self-other-mergence*

We analyzed rating and decision data using logistic general linear models (GLMs) applied first to each subject separately and then averaged resulting regression weights (beta weights) across subjects After testing whether both ability judgments in the rating data and engage/avoid choices were based on the previous performances of the appropriate players, we went on to investigate self-other-mergence (Fig.10a).

First, we tested which information subjects used to estimate self and others' abilities. Analysis of the rating data showed that *S-ability* and *O-ability* were based on the performance history of S and O, respectively. This indicates that subjects indeed used performance feedback to generate predictions about their subsequent minigame performance (Fig.10b-i,c-i). Recent performance feedback was predictive of subjects S and O ratings. However, performance that had occurred more remotely in time had much less of an impact on the subjects' ability estimates. This finding that recent and remote events have greater and lesser impacts respectively is similar to the finding that action value estimates are based more on recent rewards in reinforcement-type learning (Sutton and Barto, 1998). Therefore, to summarize the performance histories that subjects had observed, we fitted a standard reinforcement learning (RL) model individually for each subject. The median learning rate was 0.39 (SEM = 0.04) and the median beta was 0.47 (SEM = 0.05). *S-performance* and *O-performance* from the RL models were estimates of expected performance and represented recency-weighted averages of the performance feedback (not the true performance) of S and O.

Next, we examined whether *S-performance* and *O-performance* from the RL models predicted decisions to engage in cooperation/competition. We found that a better *S-performance* increased the likelihood of cooperating and competing, while a better

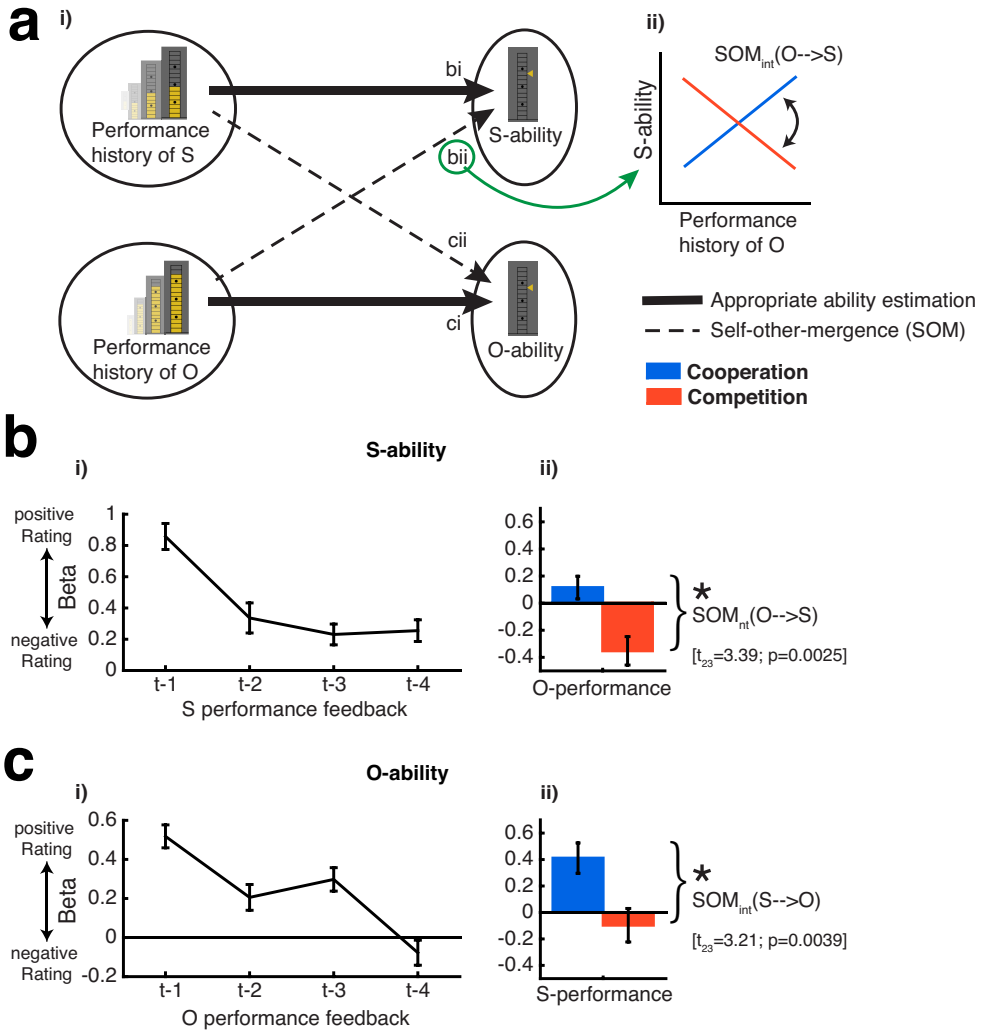


Figure 10. Rating analysis. *Blue and red indicate cooperation and competition trials in all panels.* ***(a)*** *Schematic of appropriate ability estimation and self-other-mergence (SOM) (i). Letters on arrows refer to subsequent panels showing analysis of appropriate ability estimates (b-i, c-i) and SOM effects (b-ii,c-ii). S-ability and O-ability refers to trialwise ratings of oneself (S) and relevant other (O). (ii) Illustration of a context-dependent SOM effect where O's past performance influences S-ability estimates. S-ability estimates are inflated or depressed when O's performance history has been, respectively, good or poor in cooperative contexts. However, the inverse is the case in competition: good performers decrease S-ability estimates. Positive and negative beta weights in b-ii (and analogously for O-ability in c-ii) reflect the increasing and decreasing slopes in this illustration.* ***(b, c)*** *Agent appropriate estimation of ability in panels i: S-ability and O-ability on trial t were based on recent S and O performance feedback, respectively. Panels show beta weights of a logistic GLM, applied to each subject and averaged over subjects (same for subsequent panels). Ratings were more strongly based on the recent performance feedback received by the appropriate player (i.e. S-ability ratings reflect S's past performance and O-ability ratings reflect O's past performance). SOM of performance in panels ii: We found an influence of O-performance on S-ability, controlling for the effect of S-performance. This effect reversed with social context such that the estimation of one's own ability was either inflated (in cooperation) or depressed (in competition) when paired with a high performer. Similarly, S-performance influenced estimates of O-ability, controlling for the influence of O-performance. (significant SOM_{int}; p-values for both analyses calculated in an interaction analysis from Fig.12a; same y-axes as in i). Note that different GLMs were used for panels i and ii (error bars are +/- SEM).*

O-performance led to increased cooperation and decreased competition (Fig.11a). The third factor, the threshold that subjects had to exceed on each trial not surprisingly also influenced behavior; a higher threshold led to less engagement in both cooperative and competitive contexts. As incentivized by the experimental design, subjects' engage rates were around 70% (Fig.11b) and we found a slight preference for cooperation compared to competition (t_{23}=3.82; p<0.01). This preference might be related to the experiment's payoff structure. However, it is also consistent with suggestions that a normative cooperative bias exists in social interactions promoting cooperation over selfish behavior (Boyd and Richerson, 2009). This shows that choices when to cooperate and when to compete, like the subjects' ability ratings, were strongly based on the previous performances of S and O. It also indicates that our competition/cooperate manipulation had the expected impact on subjects' behavior: subjects preferred to cooperate with high performers and to compete with low performers, as they should indeed have done in order to maximize their rewards in the experiment.

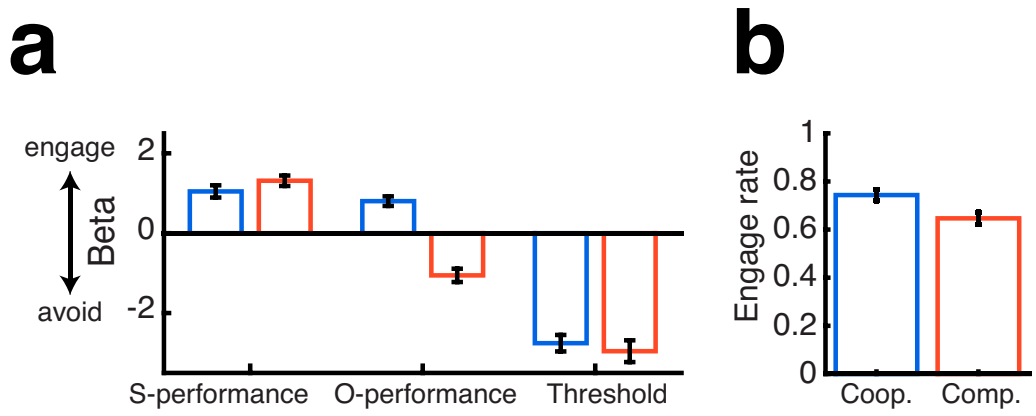


Figure 11. Choice analysis. *(**a**) S-performance, O-performance and threshold influenced decisions to cooperate or compete in a rational manner. In particular, O is weighted in reverse fashion depending on whether O is an ally in a cooperative context or an opponent in a competitive context. (**b**) Rate of engage choices for cooperative ("Coop.") and competitive ("Comp.") trials. (error bars are +/- SEM).*

Finally, we investigated whether subjects' own ability estimates were also related to O's past performance. We focused our analysis of SOM on trials where subjects chose to engage in cooperation or competition. Controlling for *S-performance*, we assessed the influence of *O-performance* on *S-ability*. To do this, we tested whether, in a GLM which contained factors indexing past performance of S, past performance of O were predictive of subjects *S-ability* ratings. Specifically, we tested whether *S-ability* judgments were indeed biased towards the performance levels of O in cooperation, but away from them in competition. We found that, indeed, during cooperation *S-ability* increased and decreased in tandem with *O-performance*: independent of subjects' own performance, subjects evaluated themselves more positively when the other player performed well and more negatively when the other performed badly. This was reversed in competition: subjects evaluated themselves more negatively when competing with a high compared to a low performer, as demonstrated by an interaction effect (SOM_{int}(O-->S), t_{23}=3.39; p=0.0025; see Fig.12a for full details and summary in Fig.10b-ii). We found *O-ability* was influenced by *S-performance* in a complementary manner (SOM_{int}(S-->O), t_{23}=3.21; p=0.0039; see Fig.12a for full details and summary in Fig.10c-ii). The two SOM effects persisted both when using estimates of S's and O's performance history that were not based on RL models and also in analyses that included the trials where subjects decided to avoid cooperation or competition (even though on such trials the relation to O might be less important, Fig.12b,c). Therefore, ability estimates for self and other were interdependent; SOM occurred both when judging oneself and the other individual.

4.3.2. *Detailed interaction GLMs*

We calculated the GLMs shown in Fig.10b-ii,c-ii (rating GLM2a, Methods) separately over compete and cooperate trials. For the GLMs presented in this section (rating GLM2b, Methods), however, the difference in effects of the observed performance

history between cooperate and compete trials was coded as an interaction effect (*S-performance x Context* and *O-performance x Context*, respectively) such that a positive interaction reflects stronger influence of *S-performance/O-performance* for cooperate compared to compete trials. We performed significance testing for the effects shown in the previous section on the GLMs presented in this section.

We found significant SOM_{int} for *S-ability* (t-test against zero; *O-performance x Context* effect: $t_{23}=3.39$; $p=0.0025$; Fig.12a) and *O-ability* (t-test against zero; *S-performance x Context*: $t_{23}=3.21$; $p=0.0039$; Fig12a). The results of these two significance tests are shown in Fig.10b-ii and Fig.10c-ii, respectively. In other words, we used the GLMs shown in Fig.12a for significance testing of the *S-performance* and *O-performance* effects binned by social context that are shown in Fig.10b-ii and Fig.10c-ii. There is also a main effect of Context on *O-ability*, suggesting that the other is evaluated better in cooperation than in competition, mirroring research demonstrating a general evaluation bias in favor of cooperating in-group members compared to competing out-group members (Brewer, 1979).

We ran two sets of control GLMs. Our first set is based on parameters that are not derived from a reinforcement learning (RL) model (Fig.12b). These GLMs used the same regressors as in Fig.12a, except that the summary terms for performance history, *S-performance* and *O-performance*, were replaced with the most recent piece of performance feedback that was received for that player (S-p(t-1) and O-p(t-1), respectively). As our main GLM, this control GLM is calculated using engage trials only. Interaction effects with social context were calculated as in Fig.12a resulting in significantly positive SOM_{int} effects for *S-ability* (*O-p(t-1) x Context*; $t_{23}=3.13$; $p=0.0047$) and *O-ability* (*S-p(t-1) x Context*; $t_{23}=2.92$; $p=0.0078$). This demonstrates that self-other-mergence effects for both *S-ability* and *O-ability* do not rely on specific parameters used by our RL model for summarizing performance histories.

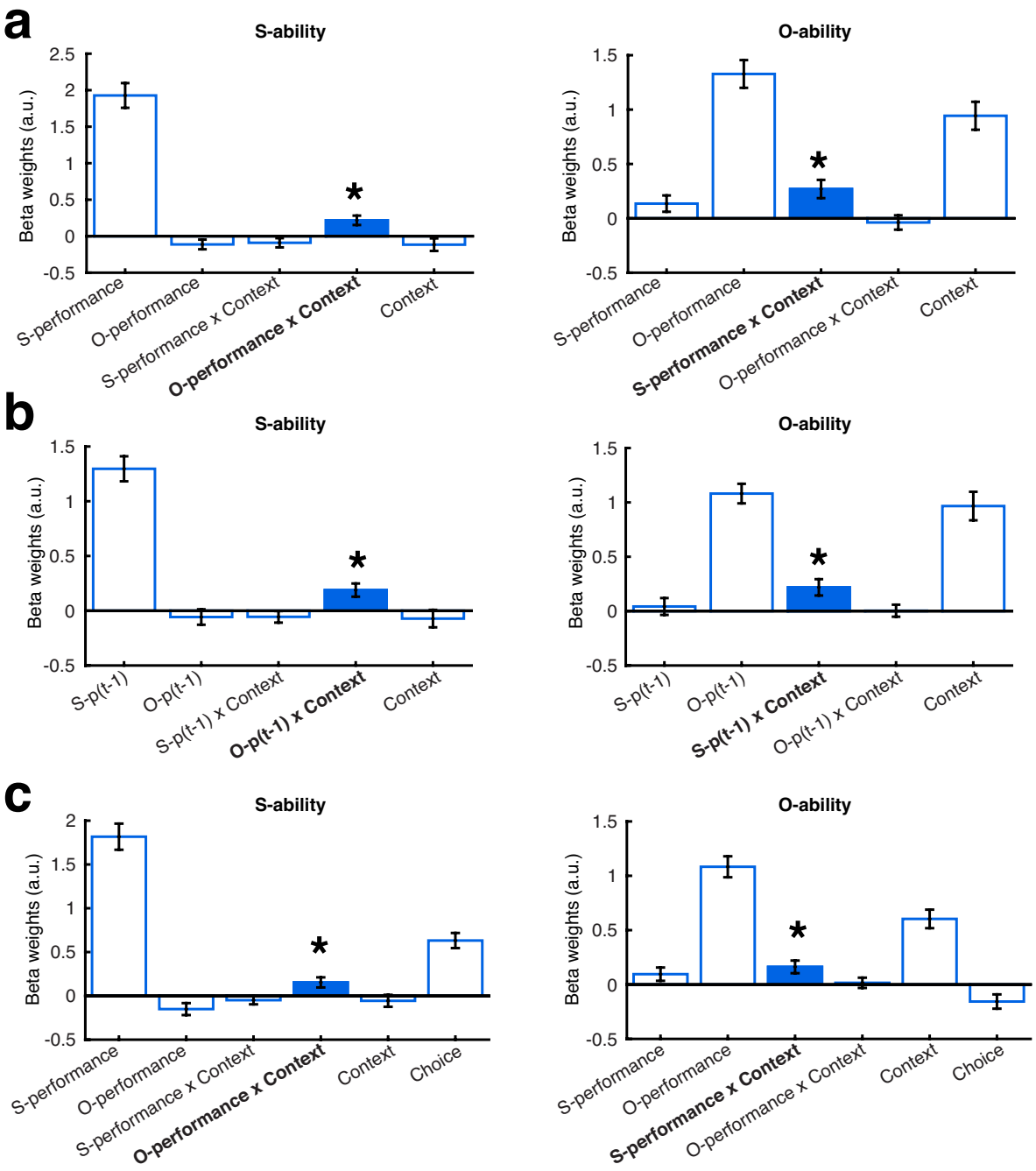


Figure 12. ***Behavioral GLMs on the self ratings (S-ability, left panels) and relevant other ratings (O-ability, right panels) for engage trials (panels a and b) and all trials (c).*** *For clarity, filled blue bars highlight self-other-mergence interaction (SOM_{int}) effects. (error bars are +/-SEM; *, $p < 0.05$)*

Our second set of control GLMs shows significant SOM_{int} when the analysis is not restricted to engage trials (as in Fig.12a,b) but also includes trials on which subjects avoided cooperation/competition. We repeated rating GLM2b (as used for Fig.12a) but

extended the analysis over both engage and avoid choice trials. Including avoid choices is problematic because O performance is irrelevant to the participants' outcomes on these trials and so any SOM effects are expected to be weaker. To account for choice, we added a binary choice regressor (normalized to a mean of zero and standard deviation of 1, positive values for engage, negative values for avoid trials). The SOM on *S-ability* (t_{23}=2.67; p=0.0138) and *O-ability* (t_{23}=2.79; p=0.0103) still remained significant.

4.3.3. *Behavioral adjustments driven by self-other-mergence*

Finally, we investigated whether self-other-mergence had an impact beyond the subjective estimates of abilities reported by subjects. We used the trial-wise change in subjects' performance, regardless of direction (*S-pChange*; Fig.13a) as an index of feedback-driven behavior strategy adjustment from one trial involving a given minigame to the next trial with the same minigame. Note that *S-pChange* refers to the change in true performance, whereas *S-performance* and *O-performance* summarize the performance feedback observed. Subjects changed their performance in the minigames (for better or worse) more after negative performance feedback for S (t_{23}=2.34, p=0.0281). However, they also changed their performance more after negative feedback for O in cooperative trials, while in competition they changed their performance more after more positive feedback for O (*O-influence on S-pChange;* t_{23}=2.25, p=0.0346; see Fig.13c for full details and summary in Fig.13b). This means that the performance feedback, displayed for the relevant other after a minigame, had an impact on how subjects performed the minigame themselves when they encountered it the next time. Therefore, subjects used the performance feedback pertaining to the other to adjust their own behavior. Like in previous analyses, significance testing was performed on the GLM using interaction terms instead of binning (Fig.13c).

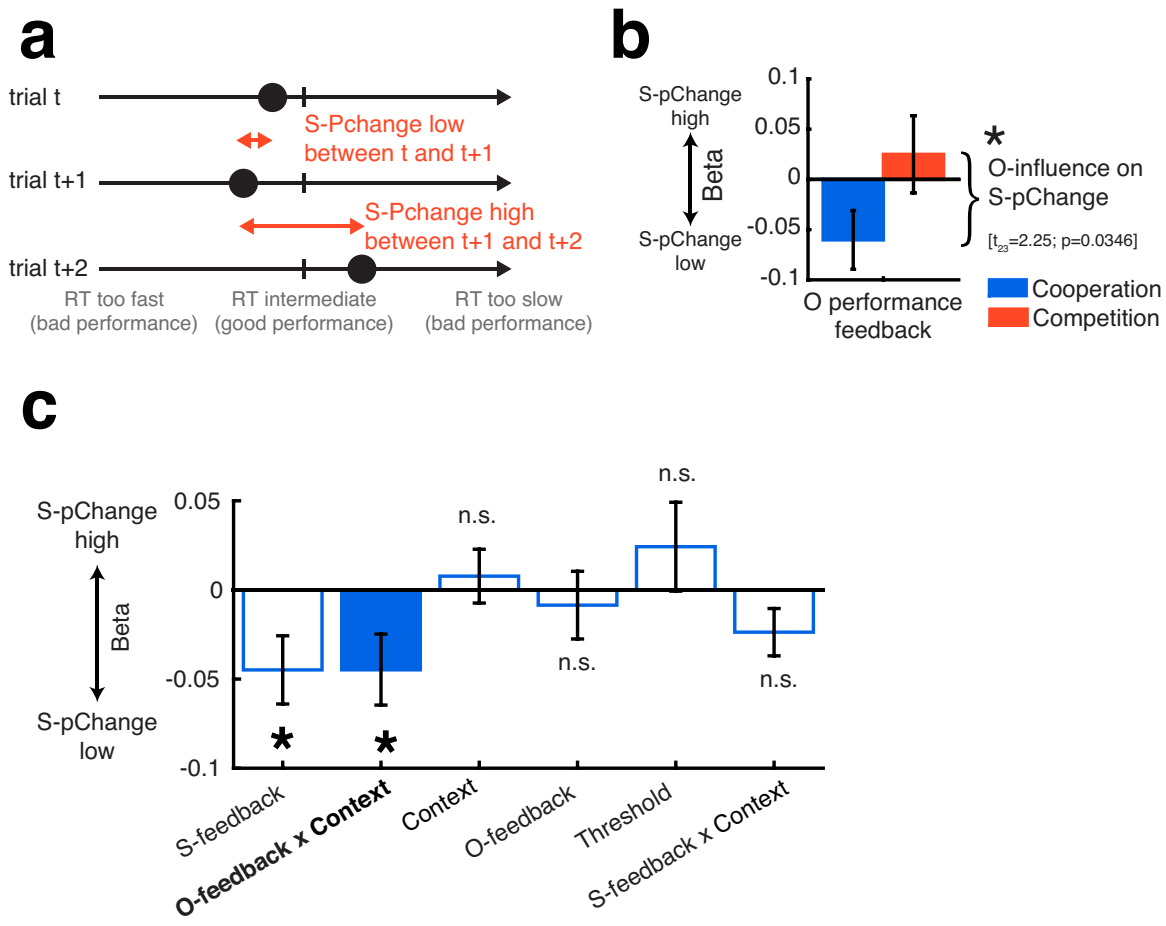


Figure 13. Self-other-mergence (SOM) in minigame performance adjustments (S-pChange). (a) *Illustration of S-pChange reflecting the absolute (i.e. unsigned) difference in true minigame performance from one trial to the next of the same minigame. S-pChange constitutes an index of how much subjects changed their true minigame performance as a function of performance feedback.* ***(b)*** *Subjects' performance adjustments were guided by their own performance feedback. S-pChange increased with more negative performance feedback. However, O performance feedback exerted an additional influence on S-pChange. It influenced S-pChange in a manner that paralleled its influence on S-ability (Fig.10; p-value from Fig.13c).* ***(c)*** *Significance testing was again performed on a GLM using interaction terms. The filled blue bar indicates the SOM related effect. S-pChange from the current to the next trial was higher when subjects received performance feedback suggesting that they performed badly (S-feedback effect, t_{23}=-2.344 ; p=0.0281). The performance feedback for the relevant other (O) also had an influence on S-pChange, depending on the social context in which it was observed. Subjects changed their performance more when the relevant Other received negative performance feedback in cooperation (just as they did for themselves) and positive performance feedback in competition (O-feedback x Context, t_{23}=-2.246 ; p=0.0346).*

Finally, in Fig.14, we show subjects' true performance data in the two minigames, illustrating that in general subjects performed the minigames well (Fig.14a) and did not have many false start trials (Fig.14b).

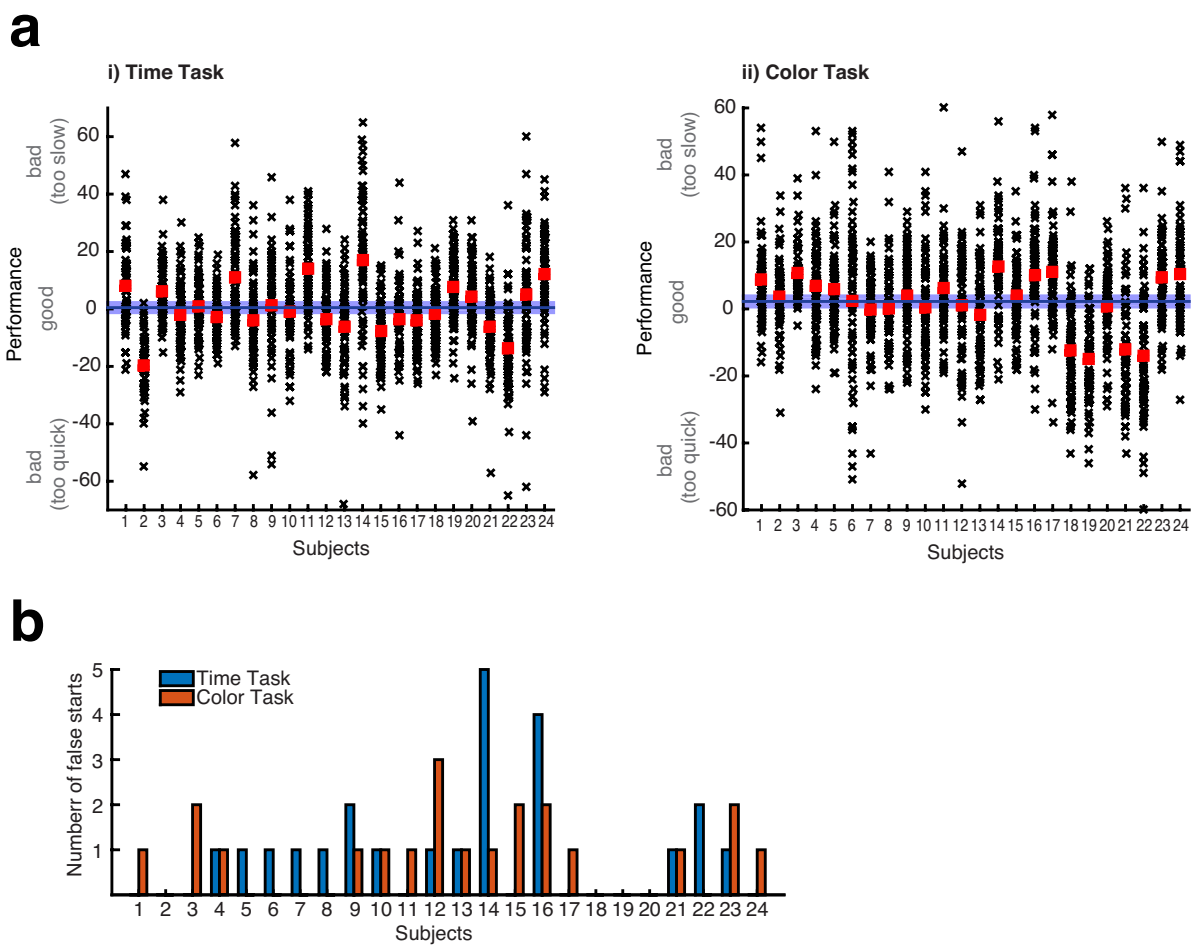


Figure 14. True performance data. (a) *Signed minigame performances plotted for all subjects for all trials (black crosses), mean signed performance per subject (red dots) and mean signed performance for the whole sample (horizontal dark blue line; shaded dark blue indicates standard error of the mean). False start performances are not shown. Note that the figure shows signed performance values, because although the absolute value is enough to indicate performance levels, the signed performances are necessary to calculate the change in performance from one trial of a minigame to the next trial of the same minigame (S-pChange). For both minigames, negative performance values indicate responses that were too quick and positive values indicate responses that were too slow. Zero indicates the optimal response time. Performances in time and color task were comparable with no differences in mean performance (t_{23}=-0.899; p=0.3778). This and the fact that the mapping of performance feedback schedules to minigames was balanced across subjects (11 received better performance feedback in the time task and 13 subjects received better performance feedback in the color task, see Fig.4) makes it unlikely that S-pChange effects were driven by performance differences in the minigames.* ***(b)*** *Number of false start trials per participant per minigame.*

4.4. Discussion

People learn about themselves from objective experience (Bandura, 1977), but their judgments are also deeply influenced by the social environment (Allport, 1924; Festinger, 1954). In our experiment, we show that both mechanisms occur when learning about the abilities of self and others. Subjects form estimates of how well they themselves and others do based on explicit performance feedback. However, these estimates are not separate, but influence each other reciprocally. Subjects systematically overestimated their abilities when cooperating with a good partner compared to a bad one and the reverse was true in competition. Also, others were estimated as more similar in ability to oneself in cooperation, but dissimilar to oneself in competition. Such interdependence could be described as "anchoring" (Tversky and Kahneman, 1974) self and other judgments to each other in cooperation, but as the reverse in competition.

Self-other-mergence might reflect difficulties that arise from tracking and assigning both self- and other-related information to the appropriate agent. In this sense the effects we report are reminiscent of failures of credit assignment to the appropriate choice during reward-guided learning (Chau et al., 2015; Thorndike, 1911, 1933; Walton et al., 2010). However, self-other-mergence may also be a side-product of relational computations. In our experiment, the cooperative and competitive contexts create a social relationship between self and other that subjects need to be aware of. This meant that expectations of success in a trial were not a result of one's own anticipated performance alone but they were also dependent on the predicted performance of another player as well. Taking into account the actions of oneself in concert with another person's actions is not a trivial matter (Tomasello et al., 2005) and it has been argued that coordinating actions between self and other often relies on implicit agreements (Misyak et al., 2014). It might be that self-other-mergence indexes the relational representations that are used in many

social situations where outcomes are the consequence of joint actions (Seo et al., 2014), possibly even if oneself is only a passive observer. From this perspective, when cooperating with a weak partner, the reduced chances of achieving a shared goal as a team might be carried over to impact negatively on one's judgment of oneself. Note however, that the strength of the direction of influence of self-other-mergence (self to other or other to self) may depend on the social situation and other constraints. Moreover, this inability to fully distinguish own competence from the competence of others might also help understanding why people who cooperate in a group often are not able to make good use of the individual knowledge they could contribute (Stasser and Titus, 1985) and, in general, why performance in groups is so hard to predict from individual performance (Faber et al., 2015).

CHAPTER 5: Neural correlates of ability estimation

5.1. Introduction

The last chapter examined human behavior in a task that required subjects to learn about their own and other people's abilities. In this chapter, fMRI data acquired during the same experiment is analyzed using the parameterizations and the computational model described in the previous chapter (Wittmann et al., 2016b).

In the field of neuroscience, we are only beginning to explore the computational and neural mechanisms that underlie how people learn about the abilities of other people (Boorman et al., 2013a), but even less is known about how we learn about our own abilities. Reward-related brain signals scale with the payoff for oneself relative to the payoff of other people (Fliessbach et al., 2007). Recently, it has been shown that one's own choice preferences can be biased towards the observed choice preferences of other people (Garvert et al., 2015; Nicolle et al., 2012) and that this depends on the identity of the other person consistent with psychological theories (Izuma and Adolphs, 2013). However, unlike choice preferences, where there is no clear right and wrong, there is often objective information available about our own abilities. For instance, the time one needs to run 100 meters ought to be well predicted by the previous occasions on which one ran 100 meters. In this sense, ability judgments can be based on objective performance feedback attributable unambiguously to ourselves.

We show that the history of an individual's performance is indeed used to estimate their ability. Moreover, we find that the performances of other people also influence individuals' assessments of their own abilities. When we are cooperating with someone who performs well our estimates of our own ability are inflated, and when we are cooperating with someone who performs poorly our ability estimates are depressed. Reciprocally, the ability of another individual is not only estimated from the other's

performance but also from one's own performance. We demonstrate that such self-other-mergence not only impacts on people's subjective evaluation of themselves and others, but even affects how they subsequently adjust their performance. We refer to ability estimation for self and others based on their respective performance history as "appropriate", while we refer to the misattribution of past self-related performance when estimating others' abilities (or other-related performance to estimate one's own ability) as self-other-mergence (SOM; see Fig.1 for an illustration). In this chapter, the neural correlates of these processes are examined.

Human subjects performed an experiment in a magnetic resonance imaging (MRI) scanner. Distinct regions in medial frontal cortex (Neubert et al., 2015) tracked the estimated abilities of self (perigenual anterior cingulate cortex; pgACC) and other (area 9)(Kelley et al., 2002; Mitchell et al., 2006). SOM increased with the strength of self and other representations in area 9; its activity predicted both how much self-judgments were

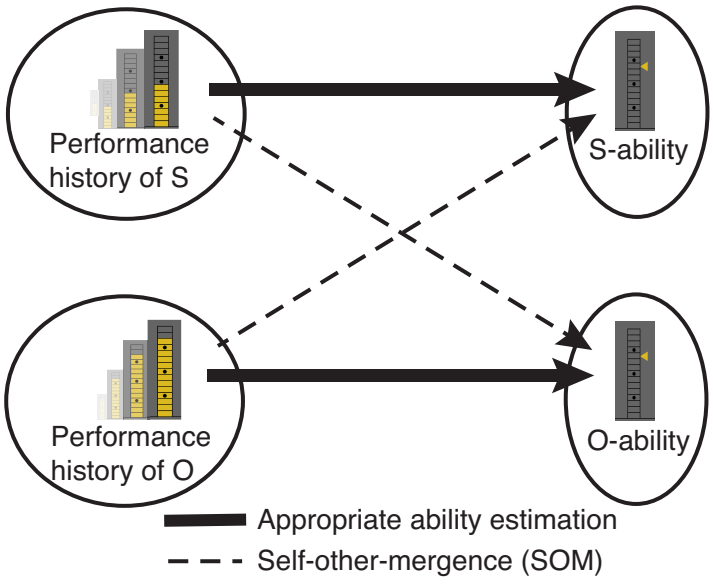


Figure 1. Schematic of appropriate ability estimation and self-other-mergence (SOM). *We refer to ability estimation for self and others based on their respective performance history as "appropriate" (solid black line), while we refer to the misattribution of past self-related performance when estimating others' abilities (or other-related performance to estimate one's own ability) as self-other-mergence (SOM; dotted line).*

related to the other player, and how much other-judgments were related to one's own performance. This suggests that area 9 does not simply represent other people's perspectives independently of our own (Amodio and Frith, 2006), but instead it represents self and others in an interdependent fashion.

5.2. Methods

5.2.1. *FMRI data acquisition and preprocessing*

Imaging data were collected on a 3 Tesla Siemens MRI scanner using a 32 channel head coil. T1-weighted structural images were acquired with the settings TR=3sec, TE=4.75mssec, TI = 1100msec, 1x1x1mm voxel size, 256x176x224 grid. Functional images were acquired using a Deichmann echo-planar imaging (EPI) sequence with TR=3s, TE=30 ms, 3x3x3mm voxel size, 87° flip angle, 30° slice angle and z-shimming to reduce signal distortions as well as dropout in medial orbitofrontal areas (Deichmann et al., 2003).

We used FMRIB's Software Library (FSL) to analyze imaging data(Smith et al., 2004). FMRI data preprocessing comprised spatial (Gaussian using full-width half maximum of 5 mm) and temporal filtering (3 dB cut-off at 100sec), motion correction with FSL's MCFLIRT and filtering of noise components after visual inspection using FSL's MELODIC. In a two-step procedure via subjects' individual structural MRI images, preprocessed functional data were nonlinearly registered to Montreal Neurological Institute (MNI) space.

5.2.2. *FMRI whole-brain analyses*

We used FSL FEAT (Smith et al., 2004) for first level analyses. FMRI data were pre-whitened with FSL FILM to account for temporal autocorrelations. Motion regressors from MCFLIRT were included as nuisance regressors of no interest. Temporal derivatives of relevant regressors were included and the model was temporally filtered before it was applied to the data. Group results on the second level were calculated using FSL FLAME 1 with outlier de-weighting and a cluster-forming threshold of $z>2.5$ and $p<0.05$.

We used a single fMRI GLM for whole brain analysis. The same design was used

for subsequent region of interest (ROI) analyses. The GLM includes RL-based regressors, which were fitted individually on all ratings and engage/avoid decisions (see last chapter for details). All parametric regressors were normalized (mean of zero, standard deviation of one). The two main phases of interest were the decision phase and the feedback phase.

The decision phase was defined as the time period from engage/avoid decision onset to the engage/avoid choice subjects made. We modeled the decision phase as a constant regressor and accompanying parametric regressors. The parametric regressors of interest comprised:

- *S-performance*
- *O-performance*
- *S-performance x Context,*
- *O-performance x Context,*
- Context (binary regressor; cooperation 1, competition -1)
- Threshold
- logRT

The two interaction terms and threshold were calculated as explained in the last chapter. logRT is the logarithm of the engage/avoid decision reaction time. The timing parameters for the parametric regressors were identical with the constant decision phase regressor, except for threshold and logRT. These regressors' onsets were delayed by one second, as the threshold was only revealed one second after engage/avoid decision onset and knowledge of the threshold was necessary to make an engage/avoid decision.

Decision phases from starter trials were excluded from the decision regressor and accounted for by a regressor of no interest. Correlations between parametric regressors are shown in Fig.2a (however, note that this correlation matrix also contains an additional

regressor that was used in the ROI version of this design, $EV_{S\text{-}Rating+O\text{-}Rating}$, but not logRT, as the latter was not a regressor of interest).

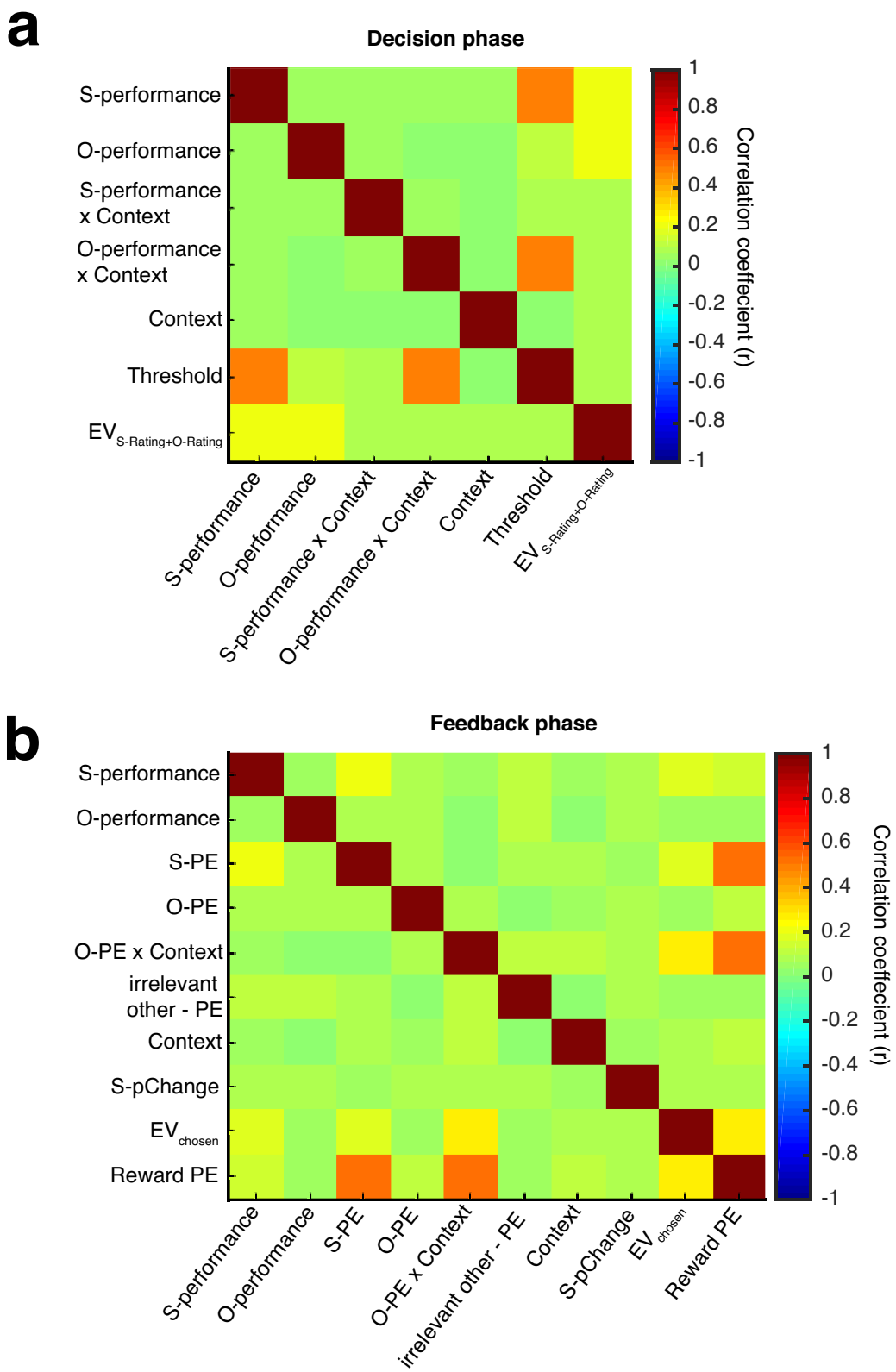


Figure 2. Correlations between regressors of interest. ***(a)*** *Correlation matrix of regressors of interest at the time of the decision over all trials.* ***(b)*** *Correlation matrix of regressors of interest at the time of feedback over all trials.*

We used two constant regressors with a duration of zero time-locked to the response of S and O rating to account for the rating events (Fig.4b shows this effect for O). In addition we used parametric regressors accompanying these constant regressors accounting for the reward expectations associated with the ratings (EV_{rating} for S and O from equation 5, see RL section in last chapter).

The feedback phase was similarly modeled as a constant regressor and parametric modulators. Note that trial feedback was chunked in three components and presented in randomized order:

I) S and O performance feedback and rating reward outcomes

II) Engage/avoid decision reward outcome

III) Irrelevant other performance feedback

Duration of the constant feedback regressor was 2.5 seconds, the time window in which the three feedback components initially appeared. Parametric regressors were modeled as stick functions (i.e. duration of zero) time-locked to the appearance of the relevant feedback component. They comprised:

- *S-performance* (I)
- *O-performance* (I)
- S-PE (Prediction error for Self) (I)
- O-PE (Prediction error for relevant other) (I)
- O-PE x Context (I)
- Prediction error - irrelevant other (III)
- Context (I)
- S-pChange (I)
- EV_{chosen} (see equation 10 in RL section in last chapter; II)
- Reward prediction error (RPE see equation 11, RL section, last chapter; II)

Roman numerals in brackets after the regressors indicate to which feedback component a regressor was time-locked. Feedback phases from starter trials, from false start trials and from trials on which no S-pChange could be calculated (see last chapter) were excluded from the feedback regressor and modeled as events of no interest. Correlations between parametric regressors are shown in Fig.2b.

In addition, the GLM contained three regressors of no interest. First, a regressor time-locked to all button presses, modeled as stick functions, to account for movement-related effects. Second, two regressors captured brain signals associated with each minigame, spanning the time period from minigame onset to response button press.

5.2.3. *FMRI region of interest analyses*

ROIs had a radius of three voxels and were centered on peak voxels of significant clusters from the whole brain GLM (Table 1). For ROI analyses, we transformed MNI to subject space and extracted the pre-processed BOLD time courses, averaged per volume. The time courses were normalized, oversampled by a factor of 20 and time-locked to decision phase onset and feedback phase onset (same onset timings as the constant regressors in the whole-brain GLM). We applied a GLM to each time point and computed one beta weight per regressor and time point, resulting in a time course of beta weights for each regressor. We extracted individual variation in signal size at the time of the group peak signal in an analysis window of 4 to 13 seconds from decision and feedback phase onset to relate brain activity to behavior. All regressors were normalized (mean of zero, standard deviation of one) for all ROI GLMs.

The ROI GLM for the decision phase contained the parametric regressors as the whole brain GLM, and, in addition contained a parametric regressor which was the sum of the EV_{rating} (equation 5, RL section in last chapter) for S and O to account for the reward expectation associated with those events:

- *S-performance*
- *O-performance*
- *S-performance x Context,*
- *O-performance x Context,*
- Context (binary regressor; cooperation 1, competition -1)
- Threshold
- logRT
- $EV_{rating\text{-}S+rating\text{-}O}$

Note that the ROI GLM for the decision phase was restricted to trials in which subjects made and engage choice rather than to avoid cooperating/competing. We did this to relate peak signal sizes from effects of interest in this analysis to behavioral beta weights from rating GLM2b (see last chapter), which was also calculated over engage trials only. The relevant beta weights were taken from the analysis in Fig.12a of the last chapter and resulting correlations with brain data are shown in Fig.3b-i,ii,iii of this chapter. In detail, these behavioral effects comprised the behavioral effect of *S-performance* on *S-ability* (Fig.3b-i), the effect of *O-performance* on *S-ability* (Fig.3b-ii; also referred to as SOM_{main}(O-->S), because it is the main effect of the relevant other's performance independent of social context) and the effect of *S-performance x Context* on *O-ability* (Fig.3b-iii).

The feedback-related ROI GLM contained the same feedback related parametric regressors as listed for the whole brain GLM:

- *S-performance*
- *O-performance*
- S-PE (Prediction error for Self)
- O-PE (Prediction error for relevant other)

- O-PE x Context
- Prediction error - irrelevant other
- Context
- S-pChange
- EV_{chosen} (see equation 10, RL section in last chapter)
- Reward prediction error (RPE; see equation 11, RL section in last chapter)

The GLM for the feedback phase was calculated over all trials. We investigated signal sizes at the time of the group peak of relevant regressors in relation to behavioral beta weights from S-pChange GLM1b (*O-feedback x Context*, last chapter), which was also calculated based on all trials. The result is shown in Fig.7b in this chapter. Note that, for visualization purposes, the behavioral effect is sign-reversed in Fig.7b.

We used a leave-one-out procedure on the group peak signal of the beta time course to do significance testing and avoid temporal selection biases. For every subject, we took the average beta time course of the relevant regressor based on the remaining 23 subjects. We identified the (positive or negative) group peak in the analysis window of 4 to 13 seconds from phase onset and then took the beta weight of the remaining subject at the time of that peak. We repeated this for all subjects. Therefore, the resulting 24 "peak" beta weights were selected independently of the time course of the subject analyzed. We assessed significance using t-tests on the resulting beta weights.

To illustrate some of the correlations between neural beta weights and behavioral beta weights (Fig.3b-ii,iii, Fig.7b), we used a median split procedure, in which our group of subjects was subdivided in two groups of 12 subjects based on their neural beta weight being low (i.e. below the median) or high (i.e. above the median). Mean and standard error of the behavioral beta weight of interest are shown in the bar plots of these figures (right hand side of Fig.3b-ii,iii, and right hand side of Fig.7b).

5.3. Results

5.3.1. *Appropriate ability estimation and self-other-mergence*

In close correspondence to the behavioral analysis presented in the last chapter, we tested, first, whether there are brain regions encoding ability estimates for self and relevant other. Having identified such regions, we went on to test whether self and other signals were related to the appropriate attribution of past performance to the relevant player (*S-performance* to S and *O-performance* to O) or whether they indicated a misattribution between self and other. For all MRI analyses, we used a single GLM.

In order to identify regions tracking performance history, we focused on the decision phase and used a GLM very similar to the one used in the behavioral choice analysis (Fig.2, Table 1). This allowed us to distinguish between brain signals related to the self, the relevant other, and the variable threshold subjects had to exceed to win points on each trial. Activity in perigenual anterior cingulate cortex (pgACC) reflected subjects' own recent performance history (Fig.3a-i); it increased and decreased depending on how well subjects had themselves performed recently (and, therefore, how likely they are to perform well on the next trial). This *S-performance* signal in pgACC did not simply reflect individual choice values; it was present even after partialling out rewards expected on a given trial (see next section and Fig.4,5). Such an analysis was possible because the experimental design meant that on any trial reward expectation was only partly a function of *S-performance* (in addition it also depended on the threshold and *O-performance*, in different ways on cooperative and competitive trials).

In contrast, area 9 tracked *O-performance* history (Fig.3a-ii), reflecting how well the relevant other is expected to perform in the minigames. Area 9 is part of the "theory of mind" network (Saxe, 2006) and is active during other-directed behavior (Sul et al., 2015).

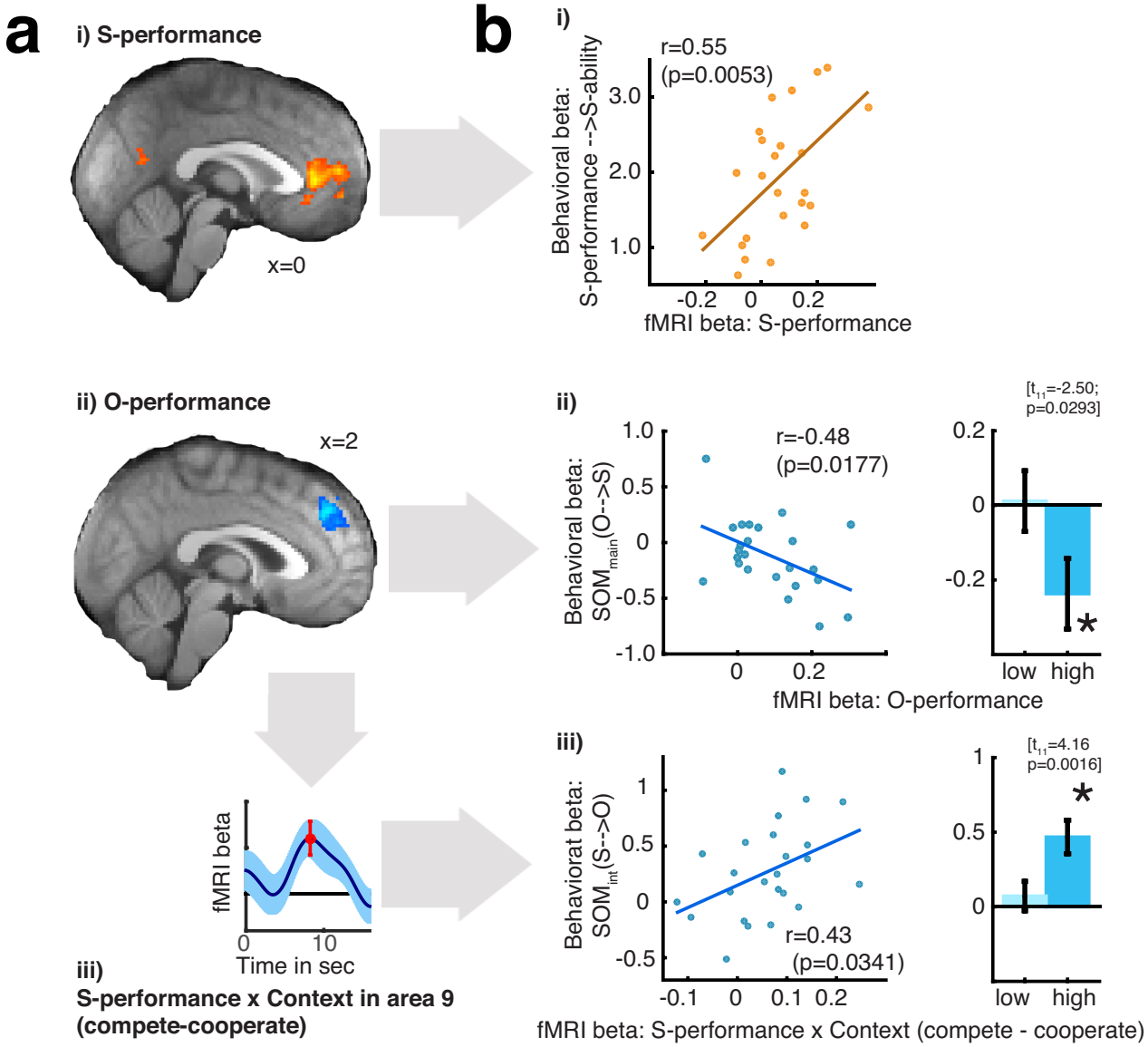


Figure 3. Representations of performance history at decision time*. Whole-brain effects family-wise error cluster-corrected, z>2.5, p<0.05.* ***(a)*** *Activation foci for S-performance in pgACC (i, yellow) and O-performance in area 9 (ii, blue). Area 9 showed more diverse social signals; a contextual S-performance signal was observed, which was stronger in competition than in cooperation (iii; beta timecourse indicates context dependent signal difference).* ***(b-i)*** *In pgACC, across-subjects variability in S-performance signal reflected the influence of S-performance on S-ability.* ***(b-ii left)*** *By contrast, variation in area 9 O-performance signal was not related to the degree subjects based O-ability on O-performance; instead it predicted the average influence of O-performance on S-ability (*SOM_{main}*(O-->S)). Note the behavioral variable plotted on ordinate corresponds to the average effect (across cooperation and competition) of O-performance on S-ability.* ***(b-ii right)*** *In other words, only subjects with high O-performance signal showed a negative effect of O-performance on S-ability (median split; same y-axis as on the left).* ***(b-iii left)*** *The strength of the contextual S-performance effect in area 9 (a-iii) predicted the degree to which O-ability was influenced by S-performance in a context-dependent manner (*SOM_{int}*(S-->O)).* ***(b-iii right)*** *Again, only subjects with higher S-performance interaction signals showed the corresponding behavioral effect of S-performance on O-ability (median split; same y-axis as on the left). (error bars are +/- SEM).*

However, the effect of *O-performance* that we found in area 9 was a basic other performance tracking signal and as such is unlikely to reflect only emotional or motivational responses to the other player that vary as a function of context. As such, the *O-performance* signal could be distinguished from activity more generally related to O. For instance, area 9 and other nodes of the "theory of mind" network were also more active at the times when subjects rated the relevant other (Fig.4b). Overall, the pattern of performance-related signals was different in area 9 compared to pgACC (see next section and Fig.5).

In all subsequent analyses we focused on the same two ROIs using the same GLM as used on the whole-brain level (although based on the same trials as in the behavioral analysis, see last chapter) using time course analysis. To relate brain signals to behavior, we employed a two-step inference process considering only variables that were related specifically to past performance. First, we assessed whether an ROI carried a significant signal for a given regressor. Second, only if this was the case, we tested whether individual variation in the brain signal was related to individual variation in behavior by correlating it with the respective behavioral beta weight from the self or other rating. We began by examining our main variables of interest: *S-performance* and *O-performance*.

We found that *S-performance* signal in pgACC predicted how much *S-performance* determined *S-ability* across subjects (Fig.3b-i; r=0.55; p=0.0053). Hence, pgACC signals reflected the use of information for estimating one's own abilities, independent of others. We found no other performance related representations of self and other in pgACC.

In contrast to pgACC, the *O-performance* signal in area 9 did not predict how much ability estimates of O were guided by *O-performance* (r=0.07; p=0.74). Instead, it predicted the degree to which subjects' *S-ability* decreased as a function of engaging with a strong other player (Fig.3b-ii; r=-0.48; p=0.0177). This means that subjects with a stronger

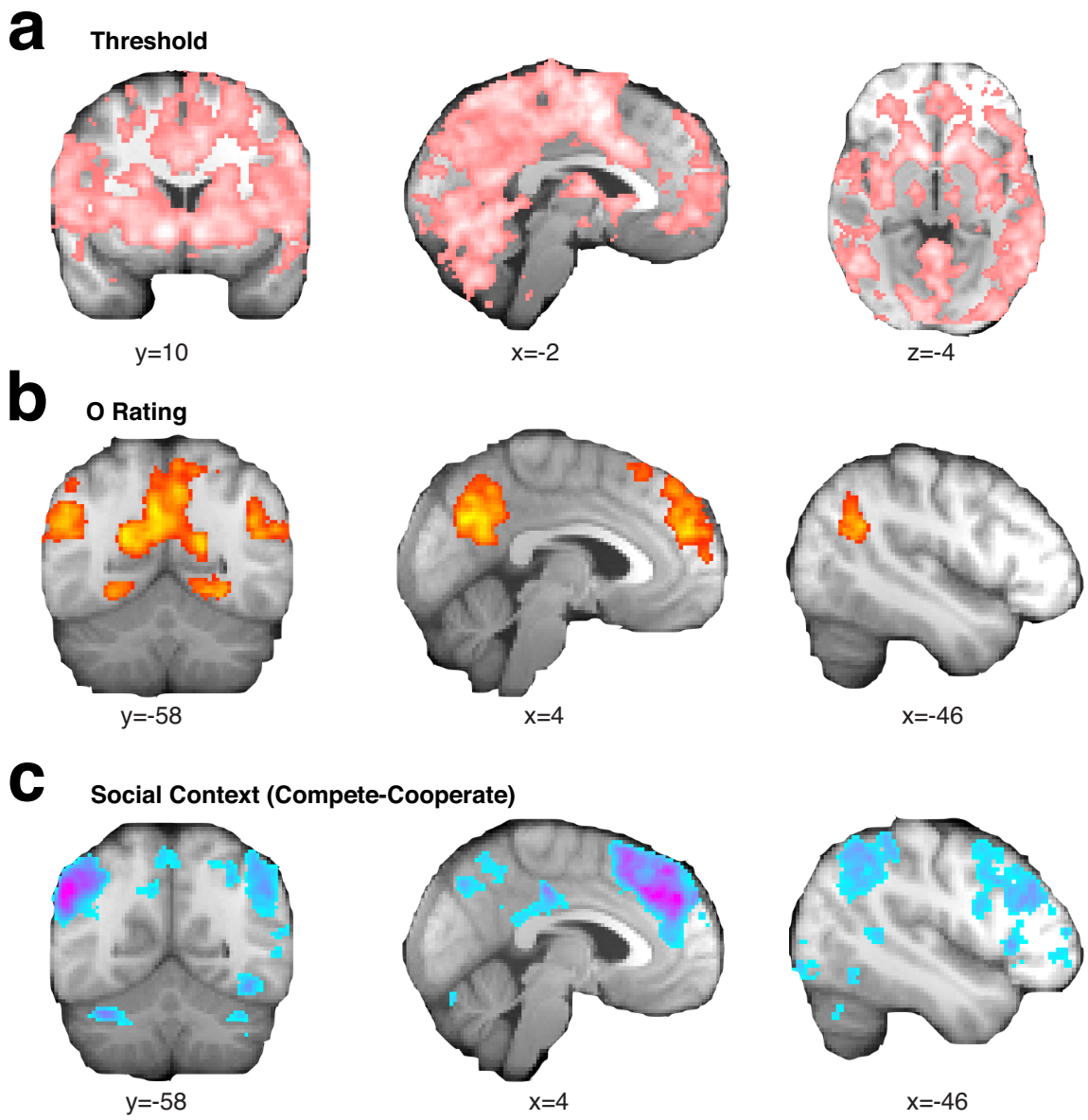


Figure 4. Other social and reward-related signals at decision time. *Whole-brain effects family-wise error cluster-corrected, z>2.5, p<0.05.* ***(a)*** *Negative effects of threshold at the time of the decision. The threshold had widespread negative effects, which in medial prefrontal cortex centered on the ventromedial prefrontal cortex (vmPFC) and not pgACC or area 9; the threshold-related activity can be considered as activity related to the decrement in reward expectation on the current trial caused by the presence of the threshold (in other words the cost of the current specific choice) rather than to S-performance or O-performance.* ***(b)*** *At the time when the rating of the relevant other was made, we found cluster-corrected positive effect (main effect of O rating, as opposed to a parametric contrast) in several brain regions typically associated with "theory of mind"-type aspects of social cognition (Saxe, 2006). These regions include dorsomedial prefrontal cortex, bilateral temporoparietal junction and precuneus.* ***(c)*** *The social context (compete-cooperate) at the time of the decision had widespread positive effects. The regions included again dorsomedial prefrontal cortex and bilateral temporoparietal junction. Although the activation was widespread, the peak of activation in dorsomedial prefrontal cortex was, just as in the case of the activations in panel b, in area 9 as identified by O-performance.*

neural representation of the relevant other's past performance in area 9 rated their own ability more negatively the better the relevant other performs. At a behavioral level this corresponded to the tendency *for O-performance* to have a negative main effect on *S-ability*. The generally negative effect of O-performance on S-ability becomes apparent when considering only subjects with the strongest *O-performance* signals in area 9 (t_{11}=-2.5; p=0.0293; Fig.3b-ii-right; median split of subjects). Therefore, area 9 activity reflected SOM in the estimation of the abilities of oneself.

To investigate whether area 9 was generally related to SOM, we examined whether such an effect was also present in the other direction: that *past S-performance* was used to estimate *O-ability*. In addition to *O-performance* signals, area 9 also carried an *S-performance* signal that was stronger in competition than in cooperation (interaction effect; $_{23}$=2.93; p=0.008; Fig.3a-iii). This was consistent with our finding that area 9 was generally more active when there was a competitive relationship between S and O than when there was a cooperative relationship (main effect of context; Fig.4c). We went on to test whether these self-related performance representations were also linked so SOM. Indeed, we found that *S-performance* influenced *O-ability* via area 9: the neural *S-performance* effect in area 9 (Fig.3a-iii) was predictive of the influence exerted by *S-performance* on *O-ability* (Fig.3b-iii; r=0.43; p=0.0341). In other words, the neural signature of *S-performance* in area 9 was different depending on context and this difference predicted the context-dependent effect of *S-performance* on *O-ability* (SOM_{int}(S-->O)).

In sum, we found that, at a first approximation, self and other signals were found in distinct brain regions (pgACC and area 9) but further investigation revealed area 9 contained more diverse social signals relating to oneself but within the social context. We found two SOM effects in area 9 that indicate that it is those subjects with stronger self and other related activity that are prone to SOM (Fig.3b-ii,iii, right panels, statistics embedded

in figure). Therefore, the stronger the representation of the other in area 9, the more the judgment of one's own ability depended also on the other's performance; analogously the strength of area 9 self-related representations predicted how much they influenced the estimation of other's abilities.

Contrast	Region	Peak Coordinates x/y/z n mm MNI space)	Z-Value
S-performance	perigenual anterior cingulate cortex (pgACC)	0 40 6	3.98
	Precuneus	-6 -64 18	3.38
O-performance	Brodman area 9	2 44 36	3.43

Table 1. *Peak coordinates of significant clusters in key fMRI contrasts.*

5.3.2. *Supplementary results related to the decision phase*

In supplementary analyses we established specifity of our correlations between behavioral and fMRI data, distinguished performance-related representations from reward expectancy and replicated neural key results for a set of independent ROIs.

First, we showed that the relationships between brain signals in pgACC and area 9 and behavior shown in Fig.3b were specific to the respective brain area. This meant that any *S-performance* signal in area 9, regardless of its significance (in fact signals were also non-significant), did not correlate with the effect of S-*performance* on *S-ability* (r=-0.27; p=0.20). Similarly, the neural signals that correlated with behavior in area 9, when measured in pgACC (again regardless of their significance), did not correlate with the behavioral variables shown in Fig.3b-ii,iii. In detail, any *O-performance* signal in pgACC did not predict the influence of *O-performance* on *S-ability* (r=-0.24; p=0.26) and *S-*

performance x Context did not predict SOM_{int}(S-->O) (r=-0.02; p=0.91). Note that Fig.5b,c show that none of these three signals was significant in the other area over all trials. Moreover, the correlation shown for pgACC was specific for the "rational" influence of *S-performance* on *S-ability* and did not occur with the "irrational" effect of *S-performance* on *O-ability* (r=0.03; p=0.88). Analogously, the behavioral correlation associated with area 9's *O-performance* signal was specific to the case of self-other-mergence (correlation of brain signal with *O-performance* effect on *O-ability*: r=0.07; p=0.74) and so was the area 9 *S-performance x Context* effect specific to the case of self-other-mergence (correlation of brain signal with *S-performance x Context* effect on *S-ability:* r=-0.04; p=0.85).

Second, we investigated the relationship between performance representations in pgACC and area 9, as well as their relationship to reward expectancy. We designed the experiment so that the correlations between EV_{chosen} and the expected performance of the players (as well as the corresponding interactions by social context) were as small as possible (Fig.5a). However expected reward is still linked conceptually to *S-performance*, *O-performance x Context* and in particular the threshold, as the reward outcome of the engage/avoid decision was a linear combination of performance feedback S, performance feedback O (positive relation in cooperation, negative relation in competition) and the threshold.

We used a repeated measures analysis of variance to formally examine whether performance related brain signals were indeed different in pgACC compared to area 9 (Nieuwenhuis et al., 2011). For this, we first used leave-one-out procedures on the whole brain level to make sure that we performed this analysis in an unbiased way. With respect to the pgACC ROI, we identified the peak voxel for *S-performance* effects based on the whole group of subjects except one left out subject. We repeated this for all subjects. Thus, we were able to conduct ROI analyses without selection biases, as each subject's brain signals did not contribute to the ROI that was selected for that subject. In the same way, we

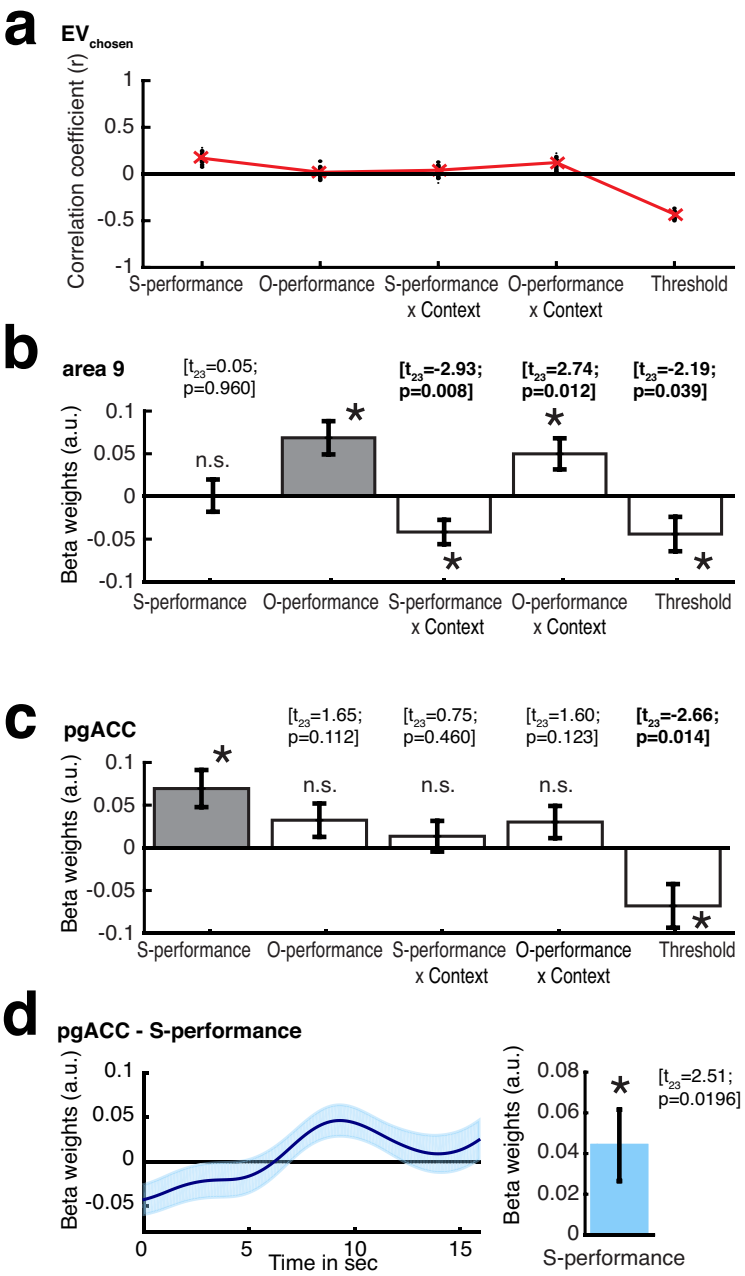


Figure 5. Signals relating to performance history and reward expectancy. *(a),* *Single subjects (small black dots) and median correlations (red line) between regressors of interest and* EV_{chosen} *(as defined in equation 10, see RL section in last chapter).* ***(b,c)*** *Effects of regressors of interest for area 9 and pgACC. Grey bars indicate contrasts used to identify the ROI. The significance stars on these effects denote the statistical significance from analyses conducted at the whole-brain level; these effects were not again tested for significance in the ROI (to avoid "double dipping"). Note that the interaction term with social context is coded in a way that positive effects mean stronger signals for cooperate compared to compete trials (interaction uses "1" for cooperation, "-1" for competition). In Fig.3, the interaction term S-performance x Context is coded the other way around in the fMRI analysis shown, so that positive signals indicate stronger responses in compete compared to cooperate trials. This was done for visualization purposes only. The stars indicating significance of those effects were derived from the whole-brain analysis.* ***(d)*** *Analysis of S-performance activity in pgACC in relation to* EV_{chosen}*. The plots show the effect of S-performance in pgACC after partialling out* EV_{chosen}*.*

identified the area 9 ROI using *O-performance*. We used a 2 [brain region] x 4 [performance signal] repeated measures ANOVA to compare the performance-related brain signals in both areas. The ANOVA included all four agent-specific signals that we investigated at the time of the decision (*S-performance, O-performance, S-performance x Context, O-performance x Context*) and compared their neural effect sizes for area 9 and pgACC. As expected, we found a highly significant interaction effect ($F_{3,69}$=6.7; p=0.0005). Main effects were absent (both p>0.08).

As *S-performance* is conceptually related to EV_{chosen} (better *S-performance* leads to higher payoff in cooperation and competition) compared to *O-performance* (better *O-performance* leads to higher payoff in cooperation but to lower payoff in competition), we went on to analyze formally if *S-performance* signals in pgACC could be reduced to EV_{chosen}. First, we used again a leave-one-out procedure to determine individual pgACC ROIs using *S-performance*. Within these ROIs, we conducted a time-course analysis using EV_{chosen} as the only regressor. We reapplied our original GLM (fMRI GLM for decision phase) to the residuals of this time course. In other words, we conducted our GLM of interest on a timecourse where the effects of EV_{chosen} had been partialled out. The effect of *S-performance* in this GLM is shown in Fig.5d as a time course analysis (left) and as a bar chart analysis (t_{23}=2.51; p=0.0196; right). In sum, *S-performance* effects in pgACC persist after partialling out EV_{chosen} and therefore cannot be explained by the expected reward of the choice alone.

Lastly, we tested the statistical solidity of our findings by examining whether our effects of interest could be replicated in two independent ROIs. A recent meta-analysis of fMRI studies investigated trait judgments about self and others (Denny et al., 2012) and identified brain regions in medial frontal cortex where self-related judgments show increased BOLD activity compared to other-related judgments (self>other) and vice versa (self<other). The two regions correspond to our regions pgACC (Fig.6a) and area 9

(Fig.6b), respectively. We replicated our effects of interest shown in Fig.3a for ROIs centered on the peak coordinates of these clusters. In detail, we again found a significant effect of *S-performance* in pgACC (t_{23}=3.61, p=0.002), and effects of *O-performance* (t_{23}=2.56; p=0.018) and *S-performance x Context* (t_{23}=-2.67; p=0.014) in area 9. Therefore, our results are not dependent on using a specific ROI location and they link up with the results of previous studies investigating similar concepts.

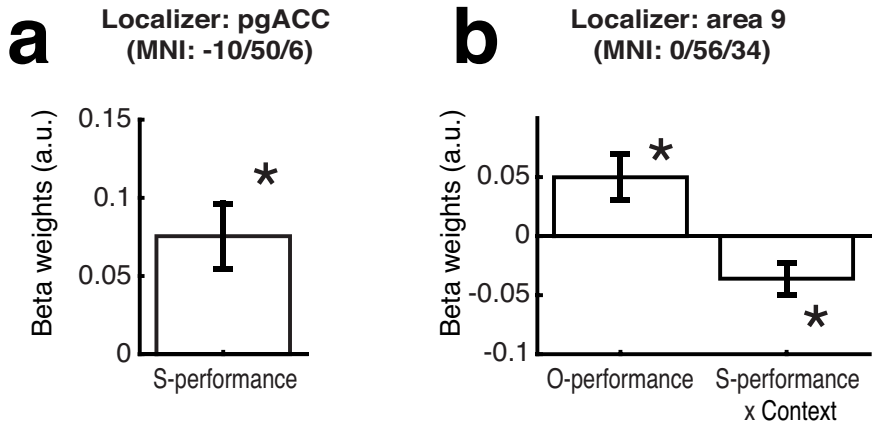


Figure 6. Supplementary control analysis for self and other related signals in pgACC and area 9. *We repeated our ROI analysis for two independent ROIs. Note that, as in Fig.5, the S-performance x Context effect is sign flipped compared to Fig.3 for ease of visualization. (error bars are +/-SEM; *, p < 0.05)*

5.3.3. *Behavioral adjustments linked to self-other-mergence in area 9*

As a last part of our investigation, we examined whether the SOM-related effects we found in the behavioral adjustment of true minigame performance (see last chapter) were mediated by area 9. We used the same area 9 ROI as before (same ROI as in Fig.3). This final part of the analyses was specifically intended as a replication of previous SOM effects.

We used the same GLM as before, focusing on the feedback phase this time. Again, we first aimed to identify activity related to the relevant other's performance in area

9 and only subsequently to relate such activity to how much subjects were biased by the other player. First we tested whether area 9 could play a role in preparing behavioral adjustments from trial to trial. We found that, indeed, in the feedback phase of the trial, activity in area 9 was correlated with the extent that subjects changed their performance from the current trial to the next trial (main effect *S-pChange*; see last chapter). Second, at the same time, we found a negative signal scaling with the size of the (signed) prediction error for the relevant other (O-PE; Fig.7a). O-PE indexed how much the observed

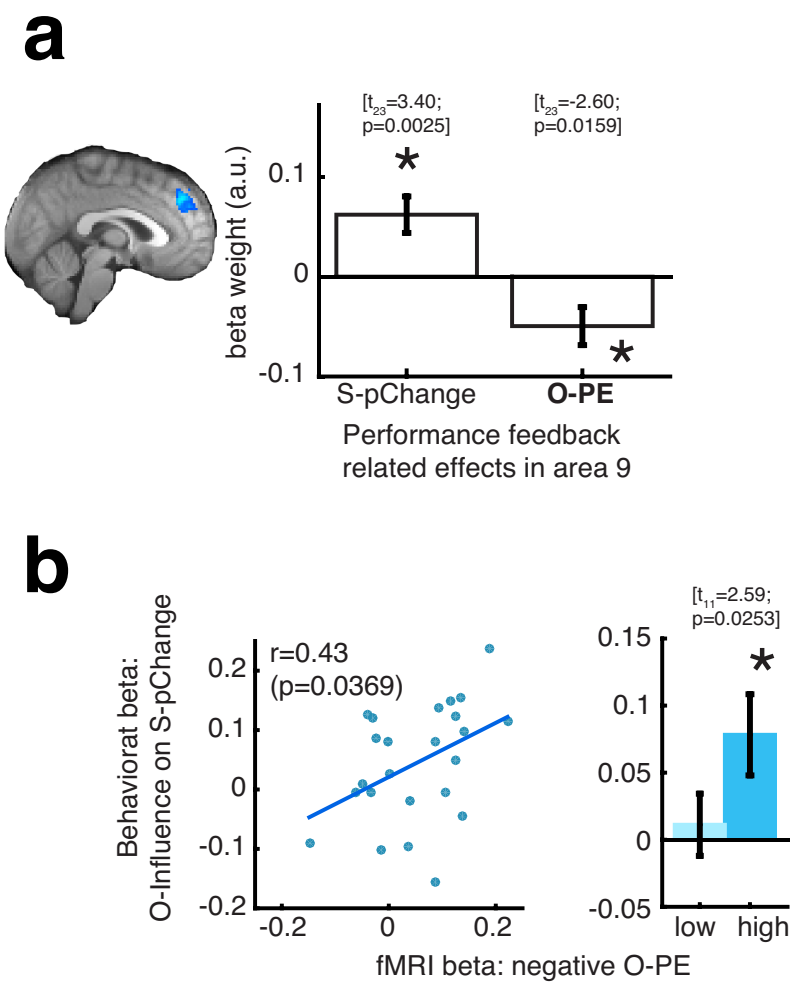


Figure 7. Self-other-mergence (SOM) in minigame performance adjustments (S-pChange). (a) *BOLD activity in area 9 in the feedback phase of trials was related to the magnitude of S-pChange that would ensue from the current to the next trial (main effect). It also scaled negatively with a prediction error for O (O-PE). ROI taken from Fig.3a-ii.* ***(b)*** *Subjects with stronger (more negative) O-PE in area 9 were more influenced by O performance feedback in their subsequent minigame performance (left). As in the earlier results, SOM occurred only in subjects with stronger area 9 signal (median split; same y-axis as on the left). (error bars are +/- SEM).*

performance feedback for O deviated from subjects' expectation of O's performance level (Suzuki et al., 2012). Third, variation in O-PE signal predicted variation in *O-influence* on *S-pChange* (Fig.7b; r=0.43, p=0.0369). This replicates our previous findings linking area 9 to self-other-mergence: the strength of other-related performance signals in area 9 predicted how much subjects' adjusted their own performance as a function of the other player's performance feedback.

5.4. Discussion

Two adjacent areas in medial prefrontal cortex, pgACC and area 9, tracked self and other performance, respectively. Area 9 moreover integrated multiple signals relating also to the self and the context and it predicted the degree to which self and other abilities were estimated in an interdependent fashion.

PgACC signals tracked subjects' own performance. This area was previously found when subjects judged whether a given trait related to themselves (Denny et al., 2012; Kelley et al., 2002) and when mentalizing about oneself and similar other people (Jenkins et al., 2008; Mitchell et al., 2006), even when controlling for the effects of likeability (Mobbs et al., 2009). In our study, pgACC activity increased and decreased according to how well subjects thought they would perform in the experiment (*S-performance* signal; Fig.3). For subjects with a stronger *S-performance* signal in pgACC, performance history more strongly governed estimates of their own ability. Activity in other brain areas has been associated with action values but our findings indicate that activity in pgACC reflects the assignment of ability, or a general value, to the self rather than the value of a particular choice (Fig.4,5). Such representations in pgACC might be used to predict whether one is capable of succeeding in a given endeavor (Bandura, 1977) and as such might be subjectively perceived as rewarding in themselves. This idea is consistent with recent findings that regions in anterior cingulate cortex are related to the physical costs monkeys are willing to endure in order to obtain rewards (Amemori and Graybiel, 2012). Moreover, such value assignments to the self (rather than value assignment to specific choices) may be altered in psychological syndromes such as depression (Murray et al., 2011). Note also that using predetermined and well-controlled performance feedback was critical for our investigation of feedback-guided ability learning as it decoupled feedback-guided learning

about oneself from introspective (meta-cognitive) estimation of one's own abilities (Bahrami et al., 2010; De Martino et al., 2013a).

In contrast to pgACC, area 9 signals were more complex. There were clear signals relating to representation of the other. For example area 9 was active at the time point that subjects made ratings of the relevant other and more specifically it tracked other-related performance estimates and prediction errors (Fig.3,7). Activity in area 9 has been found frequently in previous studies when trait judgments were made about other people (Denny et al., 2012) and mentalizing about dissimilar others (Mitchell et al., 2006). In our study there were, however, also signals relating to self and the social relationship with the other. Individual differences in area 9 activity indicated how much self/other information was used in a relational way, predicting SOM both towards the self and towards the other.

This sheds light on the function area 9 may have in social cognition. Self-other-mergence might reflect difficulties that arise from tracking and assigning both self- and other-related information to the appropriate agent. In this sense the effects we report are reminiscent of failures of credit assignment to the appropriate choice during reward-guided learning (Chau et al., 2015; Thorndike, 1911, 1933; Walton et al., 2010). However, self-other-mergence may also be a side-product of relational computations in area 9 (Fig.8). In our experiment, the cooperative and competitive contexts create a social relationship between self and other that subjects need to be aware of. This meant that expectations of success in a trial were not a result of one's own anticipated performance alone but they were also dependent on the predicted performance of another player as well. Taking into account the actions of oneself in concert with another person's actions is not a trivial matter (Tomasello et al., 2005) and it has been argued that coordinating actions between self and other often relies on implicit agreements (Misyak et al., 2014). It might be that self-other-mergence indexes the relational representations that are used in many social situations where outcomes are the consequence of joint actions (Seo et al., 2014), possibly even if

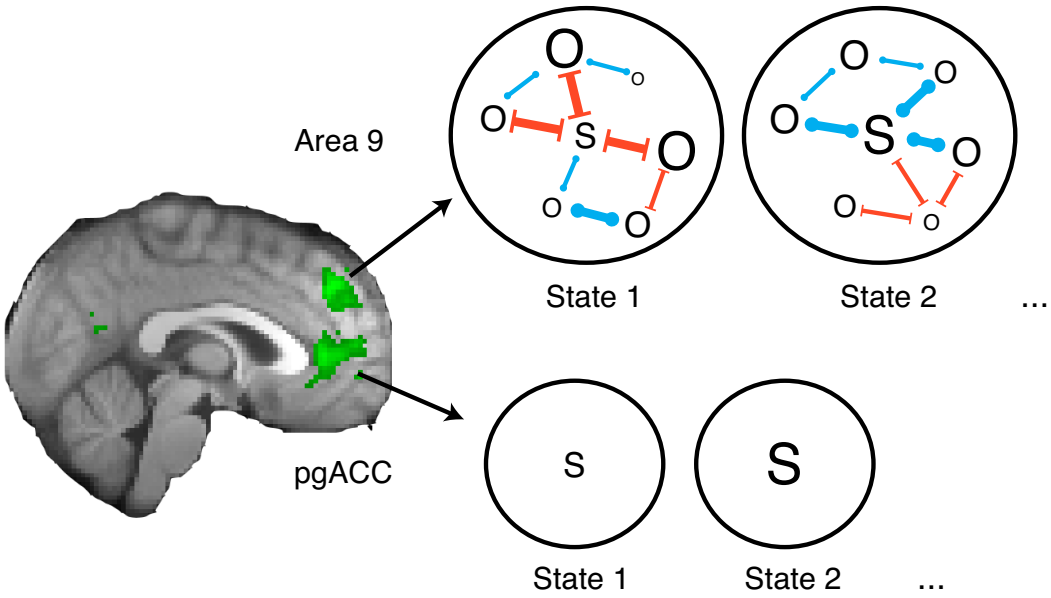


Figure 8. Self-other-mergence and interdependent self and other representations in area 9. *In our study, pgACC holds a representation of oneself that is related to observing one's own performance. It may reflect the value of one's own actions for earning future rewards or overcoming challenges. This ability estimate may be lower or higher depending on the state the individual is in (indicated by smaller and bigger "S"). In contrast, signals in dorsomedial area 9 suggest a different function. Area 9 holds a representation of the self, too, however it is embedded in the cooperative or competitive context. In addition it integrates information about the ability of others and the social relationship one has with them. A neural circuit that integrates such information might be useful to facilitate coordination between oneself and others and allow, for instance, the pursuit of shared goals. It might also be used to flexibly compute one's status within one's social network as circumstances change (blue and red connections indicate changing alliances and rivalries, varying in strength). This would enable humans and animals to assign values to self and others based on the dynamics of their social relationship - something that very vividly guides choice in the complex dominance hierarchies of monkeys. If this is the case, then self-other-mergence might accompany these neural processes in area 9 occurring as a by-product of relational self and other representations held by this brain region.*

oneself is only a passive observer. From this perspective, when cooperating with a weak partner, the reduced chances of achieving a shared goal as a team might be carried over to impact negatively on one's judgment of oneself. Note however, that the strength of the direction of influence of self-other-mergence (self to other or other to self) may depend on the social situation and other constraints. Furthermore, area 9 may integrate self, other, and relational information to compute one's own position in a social network when an

individual's status is a reflection of its alliances (Schulke et al., 2010). In monkeys, area 9 is related to dominance (Neubert et al., 2015; Noonan et al., 2014; Sallet et al., 2011) and thus, adoption of a place in a social hierarchy might reflect the operation of a relational mechanism that could be performed in area 9.

CHAPTER 6: General discussion

In this thesis, I investigated computational and neural mechanisms underlying foraging-related behavior. I examined learning and decision-making in situations that should reflect some of the challenges that animals have to overcome to survive in natural environments (Chapter 1). For instance, to maximize fitness, animals have to track the reward income of their actions over time and to also take into account that the actions of other foraging animals have a direct influence on their reward income (Stephens and Krebs, 1986). In my thesis, I looked at these types of behaviors in humans using fMRI. In particular, I investigated the neural circuitry that drives people's choices to engage and disengage with options when the profitability of their actions changes continuously over time (Charnov, 1976) and when learning about which actions oneself and others are capable of performing in multi-agent environments (Fretwell and Lucas, 1969).

I studied how people evaluate the profitability of their actions over time (Chapter 2). Of course, knowledge of how profitable alternative actions might play a key role in deciding whether to stay with the current action or whether to switch, but to make the best choices, it is important to use all information available. Tracking the value of the current action can be highly informative, particularly when the current environment is changing and alternatives are uncertain. Changes in reward rates of a single option, rather than its absolute reward rate alone, can guide choices even in rats (Crespi, 1942). I found that the value people assigned to their actions indeed comprised more than a single average reward estimate (Sutton and Barto, 1998); rewards from multiple past time points were integrated in a more complex manner. Computationally, this could be explained by reward memories with multiple time constants (Bernacchia et al., 2011) or, similarly, by keeping a memory of the reward rate change itself. The latter can be done in a simple way by estimating

average prediction errors; this can assist simple RL value estimates in guiding choice under these circumstances.

Tracking the reward income of one's actions is not only important in order to learn about the profitability of the environment, it can also inform estimates of one's own and other people's abilities (Chapter 4). Ability estimates can be used in a direct way to predict the reward outcome that the actions of oneself and others will have, without the need for more complex belief inference (Amodio and Frith, 2006; Saxe, 2006; Tomasello et al., 2005). However, the consequences of these actions may critically depend on the relationship between self and other, as others' actions can have positive or negative consequences on one's reward rate (Creel and Creel, 1995; Harper, 1982). I found that the ability estimates made for the self and the other were not kept separate. Instead people tended to judge themselves as similar to people they were cooperating with and dissimilar to people they were competing with (correspondingly they judged the other person as more similar to themselves when they were cooperating). Although this specific phenomenon was not previously reported, it resonates with a wealth of observations in social psychology that have documented how interactions between the self and the other influence how evaluations of each individual are made (Allport, 1924; Brewer, 1979; Festinger, 1954; Heider, 1958; Krueger, 2007; Mussweiler, 2003), in particular with the finding that another person's traits are estimated as more similar to one's own traits when cooperating compared to when competing with them (Toma et al., 2010).

In both experiments described in the previous paragraphs, I used simple RL models to describe how people learn the value of their actions. In both cases, however, action values comprised more than a simple average of recent reward or recent performance: the value of foraging in a patch was in addition determined by the reward rate change and estimates of self and other reciprocally influenced each other. In general, there are many cases in cognitive neuroscience where value estimates are influenced by or referenced to

additional information, for example during trinary choice (Chau et al., 2014; Louie et al., 2013; Noonan et al., 2010), credit assignment (Chau et al., 2015; Jocham et al., 2016; Walton et al., 2010) or neural range adaptation (Padoa-Schioppa, 2009). A question in these cases is often in which respect the deviation from a simple RL mechanism supports adaptive behavior. Referencing current reward rates to previous ones is an easy way to compute reward trends and doing so seems an adaptive mechanism to guide choice whenever reward trends are predictive of future values. It is unclear, however, why merging of self and other ability estimates arises. There are several possible explanations. For example, it might occur because it is computationally hard to separately track self and other's actions. According to this view failures to track one's own and the other's ability independently are the result of failures of credit assignment of an outcome to the correct agent (as opposed to the correct choice) (Akaishi et al., 2016; Thorndike, 1933). Alternative explanations might be developed in relation to the way in which agents do not just monitor the success of their own actions but instead must also compute the consequences that their own actions have on the other person and *vice versa* (Tomasello et al., 2005). Predicting whether another person's actions decrease or increase one's own reward rate might often require taking into account what the other person is capable of doing, what oneself is capable of doing and the social relationship between self and other.

Neurally, I identified several types of value representations in medial frontal cortex, for example in dACC (Chapter 3). DACC possesses neurons with a reservoir of reward time constants (Bernacchia et al., 2011; Seo and Lee, 2007) and feedback-related activity in dACC guides behavioral adjustments (Matsumoto et al., 2007; O'Reilly et al., 2013), for example by controlling the time constant with which rewards are integrated (Behrens et al., 2007b). Building on these results, I found that rewards with multiple time constants influenced BOLD activity in dACC in a way that predicted decisions to stay or to leave a foraging patch. Signals in dACC were positively related to recent reward rates and

negatively related to past reward rates, mirroring the effects that reward delivered at those times had on choice. Notably, dACC activity was decision-related rather than feedback-related and did not reflect choice difficulty (Kolling et al., 2016). However, the type of decision was one that was embedded in a series of actions with the choice being whether to continue with the current action or to switch to an alternative (Charnov, 1976). My results show that dACC in such situations can support learning of the value of the current foraging choice in a temporally detailed manner. In addition, dACC in similar settings can carry representations of the average value of the alternative foraging options (Kolling et al., 2012). Together, these findings suggest a key role for dACC in decisions to continue engaging with a current choice or to explore alternative choices in the environment.

PgACC and area 9 carried different types of action-related value representations (Chapter 5). In my study, they did not signal the value of the current foraging choice. PgACC reflected self-related ability estimates. In general, the reward rate of an action might reflect the profitability of the currently exploited environment, but it might also reflect one's own ability to perform the action. Estimates of one's own ability correspond to the reward rate of the action independent of temporary features of the environment. PgACC might thus compute a longer-term, de-contextualized value of actions. This is consistent with observations that, unlike pgACC, usually dACC and vmPFC activity reflects variables particularly relevant to the current choice (Rushworth et al., 2012), that pgACC is active during self-related trait judgments, i.e. judging how likely one is to act in a certain way (Denny et al., 2012; Kelley et al., 2002) and that pgACC signals individual attitudes towards cost-benefit decision-making (Amemori and Graybiel, 2012; Hare et al., 2011a; Kolling et al., 2012). Area 9, by contrast, represented estimates of the other's ability, but also integrated self-related ability information and contextual information. The mergence that occurred between self and other ability estimates was linked to area 9 signals in three independent analyses. The wealth of signals and the tight link between self

and other representation in area 9 is mirrored by results from single cell recordings in the macaque that show a detailed memory of past action-reward associations near this area (Donahue et al., 2013; Seo et al., 2014). These representations could enable behavior that deviates from typical RL based choice, but it could also support behavior that takes into account actions of oneself and another agent in parallel. In humans value representations in area 9 usually do not guide choices directly (Nicolle et al., 2012). Area 9 might thus provide predictions about the consequences that actions made by both oneself and by another person have on each agent. Such information may only affect decisions indirectly via other brain areas such as vmPFC and the supplementary motor and eye fields.

In sum, I have used an ecological framework to investigate learning and decision-making in humans using functional magnetic resonance imaging. I have identified computational mechanisms that can support tracking the value of actions over time and examined how people learn about which actions oneself and others are capable of performing in multi-agent environments. Different subregions of medial frontal cortex carried different types of action-related value representations that can guide decision-making: dACC tracks the value of pursuing the current foraging choice, pgACC estimates the general, de-contextualized value of an action, and area 9 integrates information about the actions of both one's own self and of other people.

Bibliography

Akaishi, R., Kolling, N., Brown, J.W., and Rushworth, M. (2016). Neural Mechanisms of Credit Assignment in a Multicue Environment. J Neurosci 36, 1096-1112.

Alexander, W.H., and Brown, J.W. (2015). Hierarchical Error Representation: A Computational Model of Anterior Cingulate and Dorsolateral Prefrontal Cortex. Neural Comput 27, 2354-2410.

Allport, F.H. (1924). The group fallacy in relation to social science. J Abnorm Soc Psychol 19, 60-73.

Amemori, K., and Graybiel, A.M. (2012). Localized microstimulation of primate pregenual cingulate cortex induces negative decision-making. Nat Neurosci 15, 776-785.

Amiez, C., Joseph, J.P., and Procyk, E. (2005). Anterior cingulate error-related activity is modulated by predicted reward. Eur J Neurosci 21, 3447-3452.

Amiez, C., and Petrides, M. (2009). Anatomical organization of the eye fields in the human and non-human primate frontal cortex. Prog Neurobiol 89, 220-230.

Amiez, C., and Petrides, M. (2014). Neuroimaging evidence of the anatomo-functional organization of the human cingulate motor areas. Cereb Cortex 24, 563-578.

Amodio, D.M., and Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7, 268-277.

Apps, M.A., Rushworth, M.F., and Chang, S.W. (2016). The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. Neuron 90, 692-707.

Badre, D., Hoffman, J., Cooney, J.W., and D'Esposito, M. (2009). Hierarchical cognitive control deficits following damage to the human frontal lobe. Nat Neurosci 12, 515-522.

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., and Frith, C.D. (2010). Optimally interacting minds. Science 329, 1081-1085.

Bandura, A. (1977). Self-efficacy: Toward a Unifying Theory of Behavioral Change. Psychol Rev 84, 191-215.

Barbas, H., Ghashghaei, H., Dombrowski, S.M., and Rempel-Clower, N.L. (1999). Medial prefrontal cortices are unified by common connections with superior temporal cortices and distinguished by input from memory-related areas in the rhesus monkey. J Comp Neurol 410, 343-367.

Basten, U., Biele, G., Heekeren, H.R., and Fiebach, C.J. (2010). How the brain integrates costs and benefits during decision making. Proc Natl Acad Sci U S A 107, 21767-21772.

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47, 129-141.

Beckmann, M., Johansen-Berg, H., and Rushworth, M.F. (2009). Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. J Neurosci 29, 1175-1190.

Behrens, T.E., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F. (2008). Associative learning of social value. Nature 456, 245-249.

Behrens, T.E., Woolrich, M.W., Walton, M.E., and Rushworth, M.F. (2007a). Learning the value of information in an uncertain world. Nat Neurosci 10, 1214-1221.

Behrens, T.E.J., Woolrich, M.W., Walton, M.E., and Rushworth, M.F.S. (2007b). Learning the value of information in an uncertain world. Nat Neurosci 10, 1214-1221.

Bernacchia, A., Seo, H., Lee, D., and Wang, X.J. (2011). A reservoir of time constants for memory traces in cortical neurons. Nat Neurosci 14, 366-372.

Blanchard, T.C., and Hayden, B.Y. (2014). Neurons in dorsal anterior cingulate cortex signal postdecisional variables in a foraging task. J Neurosci 34, 646-655.

Blanchard, T.C., Strait, C.E., and Hayden, B.Y. (2015). Ramping ensemble activity in dorsal anterior cingulate neurons during persistent commitment to a decision. J Neurophysiol 114, 2439-2449.

Bonnelle, V., Manohar, S., Behrens, T., and Husain, M. (2016). Individual Differences in Premotor Brain Systems Underlie Behavioral Apathy. Cereb Cortex 26, 807-819.

Boorman, E.D., Behrens, T.E., Woolrich, M.W., and Rushworth, M.F. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron 62, 733-743.

Boorman, E.D., O'Doherty, J.P., Adolphs, R., and Rangel, A. (2013a). The behavioral and neural mechanisms underlying the tracking of expertise. Neuron 80, 1558-1571.

Boorman, E.D., Rushworth, M.F., and Behrens, T.E. (2013b). Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. J Neurosci 33, 2242-2253.

Botvinick, M., Nystrom, L.E., Fissel, K., Carter, C.S., and Cohen, J.D. (1999). Confict monitoring versus selection-for-action in anterior cingulate cortex. Nature 402, 179-181.

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

Boyd, R., and Richerson, P.J. (2009). Culture and the evolution of human cooperation. Philos Trans R Soc Lond B Biol Sci 364, 3281-3288.

Brewer, M.B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. Psychol Bull 86, 307-324.

Bromberg-Martin, E.S., Matsumoto, M., Nakahara, H., and Hikosaka, O. (2010). Multiple timescales of memory in lateral habenula and dopamine neurons. Neuron 67, 499-510.

Buckley, M.J., Mansouri, F.A., Hoda, H., Mahboubi, M., Browning, P.G., Kwok, S.C., Phillips, A., and Tanaka, K. (2009). Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. Science 325, 52-58.

Buckner, R.L., Bandettini, P.A., O'Craven, K.M., Savoy, R.L., Petersen, S.E., Raichle, M.E., and Rosen, B.R. (1996). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. Proc Natl Acad Sci U S A 93, 14878-14883.

Buckner, R.L., and Carroll, D.C. (2007). Self-projection and the brain. Trends in cognitive sciences 11, 49-57.

Cai, X., and Padoa-Schioppa, C. (2012). Neuronal encoding of subjective value in dorsal and ventral anterior cingulate cortex. J Neurosci 32, 3791-3808.

Cai, X., and Padoa-Schioppa, C. (2014). Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. Neuron 81, 1140-1151.

Chang, S.W., Brent, L.J., Adams, G.K., Klein, J.T., Pearson, J.M., Watson, K.K., and Platt, M.L. (2013a). Neuroethology of primate social behavior. Proc Natl Acad Sci U S A 110 Suppl 2, 10387-10394.

Chang, S.W., Gariepy, J.F., and Platt, M.L. (2013b). Neuronal reference frames for social decisions in primate frontal cortex. Nat Neurosci 16, 243-250.

Charnov, E. (1976). Optimal foraging: the marginal value theorem. Theor Popul Biol 9, 129-136.

Chau, B.K., Kolling, N., Hunt, L.T., Walton, M.E., and Rushworth, M.F. (2014). A neural mechanism underlying failure of optimal choice with multiple alternatives. Nat Neurosci 17, 463-470.

Chau, B.K., Sallet, J., Papageorgiou, G.K., Noonan, M.P., Bell, A.H., Walton, M.E., and Rushworth, M.F. (2015). Contrasting Roles for Orbitofrontal Cortex and Amygdala in Credit Assignment and Learning in Macaques. Neuron 87, 1106-1118.

Cisek, P. (2012). Making decisions through a distributed consensus. Curr Opin Neurobiol 22, 927-936.

Clithero, J.A., and Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. Soc Cogn Affect Neurosci 9, 1289-1302.

Cohen, J.D., Dunbar, K., and McClelland, J.L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psychol Rev 97, 332-361.

Cole, M.W., Yeung, N., Freiwald, W.A., and Botvinick, M. (2009). Cingulate cortex: diverging data from humans and monkeys. Trends Neurosci 32, 566-574.

Collins, A.G., and Frank, M.J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci 35, 1024-1035.

Constantino, S.M., and Daw, N.D. (2015). Learning the opportunity cost of time in a patch-foraging task. Cogn Affect Behav Neurosci.

Creel, S., and Creel, N.M. (1995). Communal hunting and pack size in African wild dogs, Lycaon pictus. Anim Behav 50, 1325-1339.

Crespi, L.P. (1942). Quantitative Variation of Incentive and Performance in the White Rat. The American Journal of Psychology 55, 467-517.

Croxson, P.L., Walton, M.E., O'Reilly, J.X., Behrens, T.E.J., and Rushworth, M.F.S. (2009). Effort-based cost-benefit valuation and the human brain. J Neurosci 29, 4531-4541.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204-1215.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature 441, 876-879.

Dayan, P. (2009). Prospective and retrospective temporal difference learning. Network (Bristol, England) 20, 32-46.

De Martino, B., Fleming, S.M., Garrett, N., and Dolan, R.J. (2013a). Confidence in value-based choice. Nat Neurosci 16, 105-110.

De Martino, B., Kumaran, D., Seymour, B., and Dolan, R.J. (2006). Frames, biases, and rational decision-making in the human brain. Science 313, 684-687.

De Martino, B., O'Doherty, J.P., Ray, D., Bossaerts, P., and Camerer, C. (2013b). In the mind of the market: theory of mind biases value computation during financial bubbles. Neuron 79, 1222-1231.

Deichmann, R., Gottfried, J.A., Hutton, C., and Turner, R. (2003). Optimized EPI for fMRI studies of the orbitofrontal cortex. NeuroImage 19, 430-441.

Denny, B.T., Kober, H., Wager, T.D., and Ochsner, K.N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. J Cogn Neurosci 24, 1742-1752.

Doll, B.B., Duncan, K.D., Simon, D.A., Shohamy, D., and Daw, N.D. (2015). Model-based choices involve prospective neural activity. Nat Neurosci 18, 767-772.

Donahue, C.H., Seo, H., and Lee, D. (2013). Cortical signals for rewarded actions and strategic exploration. Neuron 80, 223-234.

Dosenbach, N.U., Fair, D.A., Miezin, F.M., Cohen, A.L., Wenger, K.K., Dosenbach, R.A., Fox, M.D., Snyder, A.Z., Vincent, J.L., Raichle, M.E.*, et al.* (2007). Distinct brain networks for adaptive and stable task control in humans. Proc Natl Acad Sci U S A 104, 11073-11078.

Dosenbach, N.U., Visscher, K.M., Palmer, E.D., Miezin, F.M., Wenger, K.K., Kang, H.C., Burgund, E.D., Grimes, A.L., Schlaggar, B.L., and Petersen, S.E. (2006). A core system for the implementation of task sets. Neuron 50, 799-812.

Ebitz, R.B., and Platt, M.L. (2015). Neuronal activity in primate dorsal anterior cingulate cortex signals task conflict and predicts adjustments in pupil-linked arousal. Neuron 85, 628-640.

Eldar, E., and Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. Nat Commun 6, 6149.

Faber, N.S., Hausser, J.A., and Kerr, N.L. (2015). Sleep Deprivation Impairs and Caffeine Enhances My Performance, but Not Always Our Performance: How Acting in a Group Can Change the Effects of Impairments and Enhancements. Pers Soc Psychol Rev.

Festinger, L. (1954). A Theory of Social Comparison Processes. Human Relations 7, 117-140.

FitzGerald, T.H.B., Seymour, B., and Dolan, R.J. (2009). The role of human orbitofrontal cortex in value comparison for incommensurable objects. J Neurosci 29, 8388-8395.

Fletcher, P.C., Happe, F., Frith, U., Baker, S.C., Dolan, R.J., Frackowiak, R.S.J., and Frith, C.D. (1995). Other Minds in the Brain - a Functional Imaging Study of Theory of Mind in Story Comprehension. Cognition 57, 109-128.

Fliessbach, K., Weber, B., Trautner, P., Dohmen, T., Sunde, U., Elger, C.E., and Falk, A. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. Science 318, 1305-1308.

Fretwell, S.D., and Lucas, H.L. (1969). On territorial behavior and other factors influencing habitat distribution in birds. Acta Biotheor.

Garvert, M.M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T.E., and Dolan, R.J. (2015). Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. Neuron 85, 418-428.

Genovesio, A., Wise, S.P., and Passingham, R.E. (2014). Prefrontal-parietal function: from foraging to foresight. Trends in cognitive sciences 18, 72-81.

Ghashghaei, H.T., and Barbas, H. (2002). Pathways for emotion: interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. Neuroscience 115, 1261-1279.

Glimcher, P.W. (2004). Decisions, uncertainty, and the brain: The science of neuroeconomics (MIT press).

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annu Rev Neurosci 30, 535-574.

Grabenhorst, F., and Rolls, E.T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. Trends in cognitive sciences 15, 56-67.

Hakimi, S., and Hare, T.A. (2015). Enhanced Neural Responses to Imagined Primary Rewards Predict Reduced Monetary Temporal Discounting. J Neurosci 35, 13103-13109.

Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci U S A 105, 6741-6746.

Hare, T.A., Camerer, C.F., and Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. Science 324, 646-648.

Hare, T.A., Hakimi, S., and Rangel, A. (2014). Activity in dlPFC and its effective connectivity to vmPFC are associated with temporal discounting. Front Neurosci 8, 50.

Hare, T.A., Malmaud, J., and Rangel, A. (2011a). Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice. J Neurosci 31, 11077-11087.

Hare, T.A., Schultz, W., Camerer, C.F., O'Doherty, J.P., and Rangel, A. (2011b). Transformation of stimulus value signals into motor commands during simple choice. Proc Natl Acad Sci U S A 108, 18120-18125.

Hare, T.T.a., O'Doherty, J., Camerer, C.C.F., Schultz, W., and Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J Neurosci 28, 5623-5630.

Harlow, J.M. (1868). Recovery from the passage of an iron bar through the head. Publications of the Massachusetts Medical Society 2, 327-246.

Harper, D. (1982). Competitive foraging in mallards:"Ideal free'ducks. Anim Behav 30, 575-584.

Hayden, B.Y., Heilbronner, S.R., Pearson, J.M., and Platt, M.L. (2011a). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. J Neurosci 31, 4178-4187.

Hayden, B.Y., Pearson, J.M., and Platt, M.L. (2009). Fictive reward signals in the anterior cingulate cortex. Science 324, 948-950.

Hayden, B.Y., Pearson, J.M., and Platt, M.L. (2011b). Neuronal basis of sequential foraging decisions in a patchy environment. Nat Neurosci 14, 933-939.

Heider, F. (1958). The Psychology of Interpersonal Relations. (John Wiley & Sons.).

Higo, T., Mars, R.B., Boorman, E.D., Buch, E.R., and Rushworth, M.F. (2011). Distributed and causal influence of frontal operculum in task control. Proc Natl Acad Sci U S A 108, 4230-4235.

Hikosaka, O. (2007). Basal ganglia mechanisms of reward-oriented eye movement. Ann N Y Acad Sci 1104, 229-249.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C.F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. Science 310, 1680-1683.

Huang, Y.Z., Edwards, M.J., Rounis, E., Bhatia, K.P., and Rothwell, J.C. (2005). Theta burst stimulation of the human motor cortex. Neuron 45, 201-206.

Hunt, L.T., Kolling, N., Soltani, A., Woolrich, M.W., Rushworth, M.F., and Behrens, T.E. (2012). Mechanisms underlying cortical activity during value-guided choice. Nat Neurosci 15, 470-476, S471-473.

Hunt, L.T., Woolrich, M.W., Rushworth, M.F., and Behrens, T.E. (2013). Trial-type dependent frames of reference for value comparison. PLoS Comput Biol 9, e1003225.

Isoda, M., and Hikosaka, O. (2007). Switching from automatic to controlled action by monkey medial frontal cortex. Nat Neurosci 10, 240-248.

Izuma, K., and Adolphs, R. (2013). Social manipulation of preference in the human brain. Neuron 78, 563-573.

Jenkins, A.C., Macrae, C.N., and Mitchell, J.P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. Proc Natl Acad Sci U S A 105, 4507-4512.

Jocham, G., Brodersen, K.H., Constantinescu, A.O., Kahn, M.C., Ianni, A.M., Walton, M.E., Rushworth, M.F., and Behrens, T.E. (2016). Reward-Guided Learning with and without Causal Attribution. Neuron 90, 177-190.

Jocham, G., Hunt, L.T., Near, J., and Behrens, T.E. (2012). A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. Nat Neurosci 15, 960-961.

Kable, J.W., and Glimcher, P.W. (2007). The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10, 1625-1633.

Kacelnik, A., and Bateson, M. (1997). Risk-sensitivity: crossroads for theories of decision-making. Trends in cognitive sciences 1, 304-309.

Karlsson, M.P., Tervo, D.G., and Karpova, A.Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. Science 338, 135-139.

Katz, L.N., Yates, J.L., Pillow, J.W., and Huk, A.C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. Nature 535, 285-288.

Kelley, W.M., Macrae, C.N., Wyland, C.L., Caglar, S., Inati, S., and Heatherton, T.F. (2002). Finding the self? An event-related fMRI study. J Cogn Neurosci 14, 785-794.

Kennerley, S.W., Behrens, T.E., and Wallis, J.D. (2011). Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nat Neurosci 14, 1581-1589.

Kennerley, S.W., Walton, M.E., Behrens, T.E.J., Buckley, M.J., and Rushworth, M.F.S. (2006). Optimal decision making and the anterior cingulate cortex. Nat Neurosci 9, 940-947.

Kerr, N.L., and Tindale, R.S. (2004). Group performance and decision making. Annu Rev Psychol 55, 623-655.

Kira, S., Yang, T., and Shadlen, M.N. (2015). A neural implementation of Wald's sequential probability ratio test. Neuron 85, 861-873.

Koechlin, E., Ody, C., and Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. Science 302, 1181-1185.

Kolling, N., Behrens, T., Wittmann, M.K., and Rushworth, M. (2016). Multiple signals in anterior cingulate cortex. Curr Opin Neurobiol 37, 36-43.

Kolling, N., Behrens, T.E., Mars, R.B., and Rushworth, M.F. (2012). Neural mechanisms of foraging. Science 336, 95-98.

Kolling, N., Wittmann, M., and Rushworth, M.F. (2014). Multiple neural mechanisms of decision making and their competition under changing risk pressure. Neuron 81, 1190-1202.

Krueger, J.I. (2007). From social projection to social behaviour. European Review of Social Psychology 18, 1-35.

Lee, D., and Seo, H. (2016). Neural Basis of Strategic Decision Making. Trends Neurosci 39, 40-48.

Lim, S.L., O'Doherty, J.P., and Rangel, A. (2011). The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. J Neurosci 31, 13214-13223.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150-157.

Louie, K., Khaw, M.W., and Glimcher, P.W. (2013). Normalization is a general neural mechanism for context-dependent decision making. Proc Natl Acad Sci U S A 110, 6139-6144.

Mackey, S., and Petrides, M. (2010). Quantitative demonstration of comparable architectonic areas within the ventromedial and lateral orbital frontal cortex in the human and the macaque monkey brains. Eur J Neurosci 32, 1940-1950.

Mackey, S., and Petrides, M. (2014). Architecture and morphology of the human ventromedial prefrontal cortex. Eur J Neurosci 40, 2777-2796.

Maier, S.U., Makwana, A.B., and Hare, T.A. (2015). Acute Stress Impairs Self-Control in Goal-Directed Choice by Altering Multiple Functional Connections within the Brain's Decision Circuits. Neuron 87, 621-631.

Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503, 78-84.

Mars, R.B., Klein, M.C., Neubert, F.X., Olivier, E., Buch, E.R., Boorman, E.D., and Rushworth, M.F. (2009). Short-latency influence of medial frontal cortex on primary motor cortex during action selection under conflict. J Neurosci 29, 6926-6931.

Matsumoto, M., Matsumoto, K., Abe, H., and Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci 10, 647-656.

Matsuzaka, Y., Akiyama, T., Tanji, J., and Mushiake, H. (2012). Neuronal activity in the primate dorsomedial prefrontal cortex contributes to strategic selection of response tactics. Proc Natl Acad Sci U S A 109, 4633-4638.

McGinty, V.B., Rangel, A., and Newsome, W.T. (2016). Orbitofrontal Cortex Value Signals Depend on Fixation Location during Free Viewing. Neuron 90, 1299-1311.

McGuire, J.T., and Kable, J.W. (2015). Medial prefrontal cortical activity reflects dynamic re-evaluation during voluntary persistence. Nat Neurosci 18, 760-766.

McNamara, J.M., Fawcett, T.W., and Houston, a.I. (2013). An Adaptive Response to Uncertainty Generates Positive and Negative Contrast Effects. Science 340, 1084-1086.

Michael, E., de Gardelle, V., and Summerfield, C. (2014). Priming by the variability of visual information. Proc Natl Acad Sci U S A 111, 7873-7878.

Milivojevic, B., Vicente-Grabovetsky, A., and Doeller, C.F. (2015). Insight reconfigures hippocampal-prefrontal memories. Curr Biol 25, 821-830.

Misyak, J.B., Melkonyan, T., Zeitoun, H., and Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. Trends in cognitive sciences 18, 512-519.

Mitchell, J.P. (2009). Inferences about mental states. Philos Trans R Soc Lond B Biol Sci 364, 1309-1316.

Mitchell, J.P., Macrae, C.N., and Banaji, M.R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron 50, 655-663.

Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A.J., Schweizer, S., Frith, C.D., and Dalgleish, T. (2009). A key role for similarity in vicarious reward. Science 324, 900.

Morecraft, R.J., Stilwell-Morecraft, K.S., Cipolloni, P.B., Ge, J., McNeal, D.W., and Pandya, D.N. (2012). Cytoarchitecture and cortical connections of the anterior cingulate and adjacent somatomotor fields in the rhesus monkey. Brain Res Bull 87, 457-497.

Murray, E.A., Wise, S.P., and Drevets, W.C. (2011). Localization of dysfunction in major depressive disorder: prefrontal cortex and amygdala. Biol Psychiatry 69, e43-54.

Murray, J.D., Bernacchia, A., Freedman, D.J., Romo, R., Wallis, J.D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., and Wang, X.J. (2014). A hierarchy of intrinsic timescales across primate cortex. Nat Neurosci 17, 1661-1663.

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. Psychol Rev 110, 472-489.

Nakamura, K., Roesch, M.R., and Olson, C.R. (2005). Neuronal activity in macaque SEF and ACC during performance of tasks involving conflict. J Neurophysiol 93, 884-908.

Nakamura, K., Santos, G.S., Matsuzaki, R., and Nakahara, H. (2012). Differential reward coding in the subdivisions of the primate caudate during an oculomotor task. J Neurosci 32, 15963-15982.

Neta, M., Miezin, F.M., Nelson, S.M., Dubis, J.W., Dosenbach, N.U., Schlaggar, B.L., and Petersen, S.E. (2015). Spatial and temporal characteristics of error-related activity in the human brain. J Neurosci 35, 253-266.

Neubert, F.X., Mars, R.B., Buch, E.R., Olivier, E., and Rushworth, M.F. (2010). Cortical and subcortical interactions during action reprogramming and their related white matter pathways. Proc Natl Acad Sci U S A 107, 13240-13245.

Neubert, F.X., Mars, R.B., Sallet, J., and Rushworth, M.F. (2015). Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. Proc Natl Acad Sci U S A 112, E2695-2704.

Neubert, F.X., Mars, R.B., Thomas, A.G., Sallet, J., and Rushworth, M.F. (2014). Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. Neuron 81, 700-713.

Nicolle, A., Klein-Flügge, M.C., Hunt, L.T., Vlaev, I., Dolan, R.J., and Behrens, T.E.J. (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. Neuron 75, 1114-1121.

Nieuwenhuis, S., Forstmann, B.U., and Wagenmakers, E.J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. Nat Neurosci 14, 1105-1107.

Nieuwenhuis, S., Holroyd, C.B., Mol, N., and Coles, M.G. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. Neurosci Biobehav Rev 28, 441-448.

Niv, Y., Daw, N.D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl) 191, 507-520.

Noonan, M.P., Sallet, J., Mars, R.B., Neubert, F.X., O'Reilly, J.X., Andersson, J.L., Mitchell, A.S., Bell, A.H., Miller, K.L., and Rushworth, M.F. (2014). A neural circuit covarying with social hierarchy in macaques. PLoS Biol 12, e1001940.

Noonan, M.P., Walton, M.E., Behrens, T.E., Sallet, J., Buckley, M.J., and Rushworth, M.F. (2010). Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. Proc Natl Acad Sci U S A 107, 20547-20552.

O'Reilly, J.X., Schuffelgen, U., Cuell, S.F., Behrens, T.E., Mars, R.B., and Rushworth, M.F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proc Natl Acad Sci U S A 110, E3660-3669.

Ogawa, S., Lee, T.M., Nayak, A.S., and Glynn, P. (1990). Oxygenation - sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. Magn Reson Med 14, 68-78.

Ongur, D., Ferry, A.T., and Price, J.L. (2003). Architectonic subdivision of the human orbital and medial prefrontal cortex. J Comp Neurol 460, 425-449.

Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. J Neurosci 29, 14004-14014.

Padoa-Schioppa, C. (2013). Neuronal origins of choice variability in economic decisions. Neuron 80, 1322-1336.

Padoa-Schioppa, C., and Assad, J.A. (2006). Neurons in the orbitofrontal cortex encode economic value. Nature 441, 223-226.

Padoa-Schioppa, C., and Assad, J.A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. Nat Neurosci 11, 95-102.

Paus, T., Tomaiuolo, F., Otaky, N., MacDonald, D., Petrides, M., Atlas, J., Morris, R., and Evans, A.C. (1996). Human cingulate and paracingulate sulci: pattern, variability, asymmetry, and probabilistic map. Cereb Cortex 6, 207-214.

Pearson, J.M., Watson, K.K., and Platt, M.L. (2014). Decision making: the neuroethological turn. Neuron 82, 950-965.

Peelen, M.V., Atkinson, A.P., and Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. J Neurosci 30, 10127-10134.

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., and Nichols, T.E. (2011). Statistical parametric mapping: the analysis of functional brain images (Academic press).

Petrides, M. (2005). Lateral prefrontal cortex: architectonic and functional organization. Philos Trans R Soc Lond B Biol Sci 360, 781-795.

Petrides, M., and Pandya, D.N. (1999). Dorsolateral prefrontal cortex: comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns. Eur J Neurosci 11, 1011-1036.

Philiastides, M.G., Biele, G., and Heekeren, H.R. (2010). A mechanistic account of value computation in the human brain. Proc Natl Acad Sci U S A 107, 9430-9435.

Picard, N., and Strick, P.L. (1996). Motor areas of the medial wall: A review of their location and functional activation. Cereb Cortex 6, 342-353.

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

Plassmann, H., O'Doherty, J., and Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J Neurosci 27, 9984-9988.

Polania, R., Moisa, M., Opitz, A., Grueschow, M., and Ruff, C.C. (2015). The precision of value-based choices depends causally on fronto-parietal phase coupling. Nat Commun 6, 8090.

Preston, A.R., and Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. Curr Biol 23, R764-773.

Price, J.L. (2007). Definition of the orbital cortex in relation to specific connections with limbic and visceral structures and other cortical regions. Ann N Y Acad Sci 1121, 54-71.

Price, J.L., and Drevets, W.C. (2010). Neurocircuitry of mood disorders. Neuropsychopharmacology 35, 192-216.

Procyk, E., Wilson, C.R., Stoll, F.M., Faraut, M.C., Petrides, M., and Amiez, C. (2014). Midcingulate Motor Map and Feedback Detection: Converging Data from Humans and Monkeys. Cereb Cortex.

Pulliam, H.R., and Caraco, T. (1984). Living in groups: is there an optimal group size. Behavioural ecology: an evolutionary approach 2, 122-147.

Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: Current research and theory 2, 64-99.

Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2004). The neural correlates of theory of mind within interpersonal interactions. NeuroImage 22, 1694-1703.

Rodriguez, C.A., Turner, B.M., Van Zandt, T., and McClure, S.M. (2015). The neural basis of value accumulation in intertemporal choice. Eur J Neurosci 42, 2179-2189.

Roitman, J.D., and Shadlen, M.N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J Neurosci 22, 9475-9489.

Rudebeck, P.H., Saunders, R.C., Prescott, A.T., Chau, L.S., and Murray, E.A. (2013). Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. Nat Neurosci 16, 1140-1145.

Rushworth, M.F., Kolling, N., Sallet, J., and Mars, R.B. (2012). Valuation and decision-making in frontal cortex: one or many serial or parallel systems? Curr Opin Neurobiol 22, 946-955.

Rushworth, M.F., Walton, M.E., Kennerley, S.W., and Bannerman, D.M. (2004). Action sets and decisions in the medial frontal cortex. Trends in cognitive sciences 8, 410-417.

Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., and Behrens, T.E. (2011). Frontal cortex and reward-guided learning and decision-making. Neuron 70, 1054-1069.

Rutledge, R.B., Skandali, N., Dayan, P., and Dolan, R.J. (2014). A computational and neural model of momentary subjective well-being. Proc Natl Acad Sci U S A 111, 12252-12257.

Ruxton, G.D., and Humphries, S. (1999). Multiple ideal free distributions of unequal competitors. Evol Ecol Res 1, 635-640.

Sallet, J., Mars, R.B., Noonan, M.P., Andersson, J.L., O'Reilly, J.X., Jbabdi, S., Croxson, P.L., Jenkinson, M., Miller, K.L., and Rushworth, M.F. (2011). Social network size affects neural circuits in macaques. Science 334, 697-700.

Sallet, J., Mars, R.B., Noonan, M.P., Neubert, F.X., Jbabdi, S., O'Reilly, J.X., Filippini, N., Thomas, A.G., and Rushworth, M.F. (2013). The organization of dorsal frontal cortex in humans and macaques. J Neurosci 33, 12255-12274.

Saxe, R. (2006). Uniquely human social cognition. Curr Opin Neurobiol 16, 235-239.

Schlag, J., and Schlag-Rey, M. (1987). Evidence for a supplementary eye field. J Neurophysiol 57, 179-200.

Scholl, J., Kolling, N., Nelissen, N., Wittmann, M.K., Harmer, C.J., and Rushworth, M.F. (2015). The Good, the Bad, and the Irrelevant: Neural Mechanisms of Learning Real and Hypothetical Rewards and Effort. J Neurosci 35, 11233-11251.

Schuck, N.W., Gaschler, R., Wenke, D., Heinzle, J., Frensch, P.A., Haynes, J.D., and Reverberi, C. (2015). Medial prefrontal cortex predicts internally driven strategy shifts. Neuron 86, 331-340.

Schulke, O., Bhagavatula, J., Vigilant, L., and Ostner, J. (2010). Social bonds enhance reproductive success in male macaques. Curr Biol 20, 2207-2210.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. Annu Rev Psychol 57, 87-115.

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

Scoville, W.B., and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. J Neurol Neurosurg Psychiatry 20, 11-21.

Seo, H., Cai, X., Donahue, C.H., and Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. Science 346, 340-343.

Seo, H., and Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J Neurosci 27, 8366-8377.

Seo, H., and Lee, D. (2009). Behavioral and neural changes after gains and losses of conditioned reinforcers. J Neurosci 29, 3627-3641.

Shenhav, A., Botvinick, M.M., and Cohen, J.D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron 79, 217-240.

Shidara, M., and Richmond, B.J. (2002). Anterior cingulate: single neuronal signals related to degree of reward expectancy. Science 296, 1709-1711.

Shima, K., and Tanji, J. (1998). Role for Cingulate Motor Area Cells in Voluntary Movement Selection Based on Reward. Science 282, 1335-1338.

Skerry, A.E., and Saxe, R. (2014). A common neural code for perceived and inferred emotion. J Neurosci 34, 15997-16008.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E.*, et al.* (2004). Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23 Suppl 1, S208-219.

Spreng, R.N., Mar, R.A., and Kim, A.S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. J Cogn Neurosci 21, 489-510.

Stasser, G., and Titus, W. (1985). Pooling of Unshared Information in Group Decision-Making - Biased Information Sampling during Discussion. J Pers Soc Psychol 48, 1467-1478.

Stephens, D.W., and Krebs, J.R. (1986). Foraging theory (Princeton, NJ: Princeton University Press).

Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. Neuron 78, 364-375.

Stoll, F.M., Fontanier, V., and Procyk, E. (2016). Specific frontal neural dynamics contribute to decisions to check. Nat Commun 7, 11990.

Strait, C.E., Blanchard, T.C., and Hayden, B.Y. (2014). Reward value comparison via mutual inhibition in ventromedial prefrontal cortex. Neuron 82, 1357-1366.

Sugrue, L.P., Corrado, G.S., and Newsome, W.T. (2004). Matching behavior and the representation of value in the parietal cortex. Science 304, 1782-1787.

Sul, S., Tobler, P.N., Hein, G., Leiberg, S., Jung, D., Fehr, E., and Kim, H. (2015). Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. Proc Natl Acad Sci U S A 112, 7851-7856.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement learning: An introduction, Vol 28 (Cambridge, Massachusetts: MIT Press).

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J.L., Ichinohe, N., Haruno, M., Cheng, K., and Nakahara, H. (2012). Learning to simulate others' decisions. Neuron 74, 1125-1137.

Suzuki, S., Jensen, E.L., Bossaerts, P., and O'Doherty, J.P. (2016). Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. Proc Natl Acad Sci U S A 113, 3755-3760.

Thorndike, E.L. (1911). Animal Intelligence; Experimental Studies (New York: Macmillan).

Thorndike, E.L. (1933). A Proof of the Law of Effect. Science 77, 173-175.

Tom, S.M., Fox, C.R., Trepel, C., and Poldrack, R.A. (2007). The neural basis of loss aversion in decision-making under risk. Science 315, 515-518.

Toma, C., Yzerbyt, V., and Corneille, O. (2010). Anticipated cooperation vs. competition moderates interpersonal projection. J Exp Soc Psychol 46, 375-381.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behav Brain Sci 28, 675-691; discussion 691-735.

Tversky, A., and Kahneman, D. (1974). Judgment under Uncertainty - Heuristics and Biases. Science 185, 1124-1131.

Ullsperger, M., and von Cramon, D.Y. (2001). Subprocesses of performance monitoring: a dissociation of error processing and response competition revealed by event-related fMRI and ERPs. NeuroImage 14, 1387-1401.

Vogt, B.A., and Pandya, D.N. (1987). Cingulate cortex of the rhesus monkey: II. Cortical afferents. J Comp Neurol 262, 271-289.

Vogt, B.A., and Paxinos, G. (2014). Cytoarchitecture of mouse and rat cingulate cortex with human homologies. Brain Struct Funct 219, 185-192.

Wallis, J.D. (2012). Cross-species studies of orbitofrontal cortex and value-based decision-making. Nat Neurosci 15, 13-19.

Walter, H., Adenzato, M., Ciaramidaro, A., Enrici, I., Pia, L., and Bara, B.G. (2004). Understanding intentions in social interaction: the role of the anterior paracingulate cortex. J Cogn Neurosci 16, 1854-1863.

Walton, M.E.M., Behrens, T.E.J.T., Buckley, M.M.J., Rudebeck, P.H., Matthew, F.S., and Rushworth, M.F.S. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron 65, 927-939.

Wang, F., Zhu, J., Zhu, H., Zhang, Q., Lin, Z., and Hu, H. (2011). Bidirectional control of social hierarchy by synaptic efficacy in medial prefrontal cortex. Science 334, 693-697.

Wang, X.J. (2008). Decision making in recurrent neuronal circuits. Neuron 60, 215-234.

Wiens, J.A. (1976). Population responses to patchy environments. Annu Rev Ecol Syst, 81-120.

Wise, S.P. (2008). Forward frontal fields: phylogeny and fundamental function. Trends Neurosci 31, 599-608.

Wittmann, M.K., Kolling, N., Akaishi, R., Chau, B.K., Brown, J.W., Nelissen, N., and Rushworth, M.F. (2016a). Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex. Nat Commun 7, 12327.

Wittmann, M.K., Kolling, N., Faber, N.S., Scholl, J., Nelissen, N., and Rushworth, M.F. (2016b). Self-Other Mergence in the Frontal Cortex during Cooperation and Competition. Neuron 91, 482-493.

Xie, J., and Padoa-Schioppa, C. (2016). Neuronal remapping and circuit persistence in economic decisions. Nat Neurosci 19, 855-861.

Yeterian, E.H., Pandya, D.N., Tomaiuolo, F., and Petrides, M. (2012). The cortical connectivity of the prefrontal cortex in the monkey brain. Cortex 48, 58-81.

Zink, C.F., Tong, Y., Chen, Q., Bassett, D.S., Stein, J.L., and Meyer-Lindenberg, A. (2008). Know your place: neural processing of social hierarchy in humans. Neuron 58, 273-283.