Dissertation der
Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universität München

Using Bayesian modelling to uncover the behavioural and neural mechanisms of social learning and decision-making in healthy controls & psychiatric disorders

Graduate School of
Systemic Neurosciences
LMU Munich

Submitted by
Lara Raphaela Sophie Henco
13th of March 2020

Supervisor
Prof. Dr. Leonhard Schilbach
Independent Max Planck Research Group for Social Neuroscience
Max Planck Institute of Psychiatry, Munich

First Reviewer
Prof. Dr. Leonhard Schilbach

Second Reviewer
Prof. Dr. Kathrin Koch

External Reviewer
Prof. Dr. Philipp Sterzer

Date of Defense
29.05.2020

Contents

List of Abbreviations

ACC	Anterior cingulate cortex
ACCs	Sulcus of the anterior cingulate cortex
ACCg	Gyrus of the anterior cingulate cortex
ASD	Autism Spectrum Disorder
BOLD	Blood oxygen level dependent
BPD	Borderline personality disorder
dACC	Dorsal anterior cingulate cortex
fMRI	Functional magnetic resonance imaging
HC	Healthy Controls
HGF	Hierarchical Gaussian filter
MDD	Major depressive disorder
MTG	Middle temporal gyrus
dlPFC	Dorsolateral prefrontal cortex
dmPFC	Dorsomedial prefrontal cortex
OFC	Orbitofrontal cortex
vSTR	Ventral striatum
VTA	Ventral tegmental area
SCZ	Schizophrenia
SN	Substantia nigra
TPJ	Temporoparietal junction
ToM	Theory of Mind
RMET	Reading the Minds in the Eyes Test

1 Introduction

1.1 Computational Psychiatry

Psychiatric disorders are characterized by aberrant cognitive, emotional and social functions. However, the understanding of these abnormalities, their underlying mechanisms and how they give rise to psychopathology remain poorly understood (Friston, Stephan, Montague, & Dolan, 2014; Huys, Maia, & Frank, 2016; Montague, Dolan, Friston, & Dayan, 2012; Stephan & Mathys, 2014; Wang & Krystal, 2014). Consequently, psychiatry remains a medical discipline that draws upon consensus-based diagnostic instruments, namely the DSM-IV (American Psychiatric Association, 2013) and ICD-10 (World Health Organization, 1992), which rely on the categorisation of patients' verbal reports and behavioural observations.

In light of relatively low rates of recovery from psychiatric disorders, it is assumed that biomarkers (i.e. measurable indicators that reflect biological dysfunctions) could help to improve diagnostic procedures and the prediction of treatment success. However, despite longstanding efforts, no validated biomarkers exist for the vast majority of psychiatric disorders.

Computational psychiatry is a burgeoning field that aims to bridge the gap between phenomenological subjective experience and the underlying neurocognitive mechanisms and therefore may be highly relevant to the development of novel biomarkers (Paulus, Huys, & Maia, 2016). The theory-driven approaches of computational psychiatry mainly employ algorithmic models to explain behaviour and the underlying neural processes (Lis & Kirsch, 2016; Montague et al., 2012). In particular, the aim of this approach is to understand how the brain computes beliefs and how they guide optimal decision-making. Consequently, (sub)-optimal choices may rely on (aberrant) belief computations, both of which are thought to constitute central aspects in psychiatric disorders (Friston et al., 2014). Importantly, abnormal decision-making and maladaptive beliefs about the social environment may be of particular relevance for psychiatric disorders, which have been construed as disorders of social cognition and interaction (Schilbach, 2016). For this reason, the present project adopted a computational approach to investigate learning and decision-making within a social context (Lis & Kirsch, 2016; Schilbach, 2015).

1.2 Social cognition in psychiatry

1.2.1 Mentalization as transdiagnostic impairment

Previously, psychiatric disorders have been construed as disorders of social interactions that affect the reciprocal exchange between two or more individuals (Schilbach, 2015). These impairments are thought to emerge from dysfunctional beliefs about the self and others that evolve as a result of aberrant mentalization (Frith & Frith, 2006; Schilbach, 2016).

Mentalization describes the ability to understand one's own and other people's behaviours in terms of their underlying mental states such as intentions or feelings (Fonagy, Luyten, & Bateman, 2015; Frith & Frith, 2006). In social interactions, mental states are usually concealed, which is why we need to use external, overt signals to make attributions of the mental states in order to predict the actions of others (Diaconescu, Hauke, & Borgwardt, 2019). This ability is crucial for guiding behaviour in social interactions (Domes, Schulze, & Herpertz, 2009) and impaired mentalization may therefore give rise to altered beliefs and misinterpretations that lie at the core of social interaction problems in various psychiatric disorders (Schilbach, 2016).

1.2.2 Methodological advances in social neuroscience

Different paradigms have been used to probe mentalization, such as the Reading the Minds in the Eyes Test (RMET) (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) and the Strange Stories and Strange Cartoons Task (Happé, 1994). In the RMET, participants are asked to attribute mental states from photographed eyes and in the strange stories tasks, participants are asked to make inferences on the underlying intentions of the stories characters.

Traditional ToM tasks adopt a third-person perspective that involves interpreting other peoples' behaviour via observation. However, these tasks fall short of the highly dynamic and reciprocal nature of social interactions in everyday life (Schilbach et al., 2013). The rise of second-person neuroscience has therefore prompted methodological and technical developments making tasks more ecologically valid (Redcay & Schilbach, 2019; Schilbach et al., 2013). An important contribution in this endeavour came from economic exchange games (Lis & Kirsch, 2016; Montague et al., 2012; Redcay & Schilbach, 2019), in which social interactions are simulated in a more realistic, yet controlled, manner. In these tasks, participants are usually asked to play with real or alleged partners in order to maximise a profit (Lis & Kirsch, 2016; Robson, Repetto, Gountouna, & Nicodemus, 2020) requiring them to engage in mentalization (Frith & Singer, 2008). The well-controlled laboratory setting allows for the application of computational models to mechanistically explain the observed behaviour, which thus gives insight into the underlying mechanisms of (social) behaviour, providing algorithmic descriptions of feedback processing (i.e. learning) and decision-making. These computational models can be based in reinforcement learning theory or Bayesian algorithms to whose description we now turn.

1.3 Computational Models of learning

1.3.1 Reinforcement learning

Computational models of learning and decision-making propose how we should optimally integrate previous beliefs and newly observed data, e.g. feedback, in order to optimise goal-directed behaviour (Dayan & Daw, 2008). They provide putative algorithmic descriptions of how the values of different options are computed and chosen from in order to maximise outcomes. The most influential reinforcement learning algorithm (Rescorla & Wagner, 1972) computes the difference between the predicted $p^{(k)}$ and actual outcome $r^{(k)}$, i.e. prediction errors $\delta^{(k)}$, that is used to update expectations about a particular state of the world. This prediction error is the driving force of learning. For instance, in an experiment in which stimulus-reward associations are learned during multiple trials (k), the predicted value of this stimulus $p^{(k+1)}$ is a function of the current prediction $p^{(k)}$ and the prediction error $\delta^{(k)}$ weighted by a learning rate α.

$$\delta^{(k)} = r^{(k)} - p^{(k)} \qquad \text{(Equation 1)}$$

$$p^{(k+1)} = p^{(k)} + \alpha * \delta^{(k)} \qquad \text{(Equation 2)}$$

The prediction error δ is positive when the received reward is higher than predicted, and negative when it is smaller than predicted. The learning rate scales the impact of the prediction error on the belief update and accounts for the *speed* of learning, i.e. how strongly new feedback is integrated. A high learning rate would imply fast changes in predictions in light of the most recent information, a slow learning rate would imply that prediction errors do not have a strong impact on future predictions. Reinforcement learning models have contributed a great deal to our understanding of the neurobiological underpinnings of reward learning but also of social learning, i.e. learning about or from others (Lockwood & Klein-Flügge, 2019) (cf. section 1.5). However, there are a number of limitations of reinforcement learning models (Gershman, 2015; Mathys, Daunizeau, Friston, & Stephan, 2011). One major limitation is that the predicted value of a stimulus is represented as a *point estimate* rather than a probability distribution, which not only entails a prediction but also its uncertainty. Another, related limitation is that the learning rate can only vary between individuals and contexts but not within individuals. More specifically, learning rates should be adjusted depending on the certainty of the environment and the certainty of previous experiences (e.g. Behrens, Woolrich, Walton, & Rushworth, 2007). The next section presents how Bayesian learning models overcome these limitations.

1.3.2 Bayesian inference

The Bayesian brain hypothesis conceptualises our brain as an inference machine that holds internal mental representations or beliefs about states of the world, that guide our perception in the form of top down predictions (Friston et al., 2014). This internal model is also called a generative model of sensory inputs because it describes a probabilistic mapping of the hidden, latent states to the sensory signals that are observable, i.e. the probability of the data Y, given states X P(Y|X), that describe how the hidden parameters generate the sensory inputs (*likelihood)*. In the generative model, the *likelihood* of the data or sensory evidence is combined with a-priori beliefs called *priors* P(X) to infer the most likely environmental cause (X) that generated the data Y P(X|Y), which is termed *posterior.* The process by the which prior beliefs change to posterior beliefs is inference and can be performed by means of the Bayes' rule.

In Bayesian inference prior and posterior beliefs are represented as probability distributions with means and variances (inverse precisions) and the extent to which posterior beliefs change in light of new observations depends on a delicate balance between the precision of the sensory data and the precision of the prior belief (Equation 3 and Figure 1). Consequently, the belief update in Bayesian learning models resemble the equation of reinforcement learning models introduced in 1.3.1 with the difference that the learning rate is defined by the ratio between the confidence or precision of the sensory data and the precision of the prior belief (Equation 3). These are termed precision weights or dynamic learning rates because prediction errors are not always weighted to the same extent. Instead, surprising events are given more weight when the precision of the prior belief is lower relative to the precision of the sensory data (Mathys et al., 2014; Mathys et al., 2011). Crucially, Bayesian learning models can instantiate belief updating across different hierarchical levels, in which higher levels represent beliefs about more abstract features of the world.

$$\Delta\text{belief} \propto \frac{\text{precision}_{likelihood}}{\text{precision}_{prior\ belief}} \times \text{prediction error} \quad \text{(Equation 3)}$$

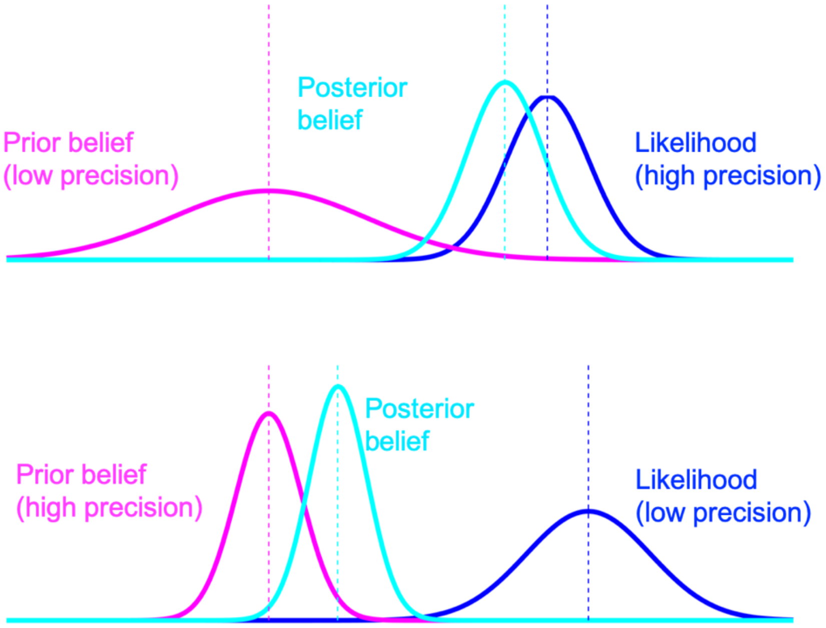


Figure 1. Bayesian belief update. This figure illustrates the principles of Bayesian inference whereby a prior belief (prediction) is combined with the likelihood (e.g. sensory data) to generate a posterior belief (updated prediction). All come in the forms of Gaussian probability distributions with means and variances (inverse precisions). The discrepancy between the prior belief and the likelihood is the prediction error. The extent to which this error is used to update the prior belief depends on the variance of the prior and the likelihood. When prior beliefs with decreased precisions meet sensory data with increased precision, the belief update is biased towards the more reliable sensory data (depicted in upper panel). If the sensory data has a decreased precision relative to the prior precision (as depicted in the lower panel), the posterior is dominated by the prior belief. The figure was created using Matlab R2017a.

1.3.3 Mentalization as special case of inference

As discussed above, the delicate balance between the precision of the data and precision of the prior belief is a prerequisite for adaptive inference (Haker, Schneebeli, & Stephan, 2016). In turn, it has been argued that an imbalance in those quantities could give rise to maladaptive beliefs in psychopathology (Mathys, 2016; Stephan & Mathys, 2014). Just as we need to infer the causes of our sensations, we need to infer the causes of the actions and emotions of others and this happens, to a large extent automatically outside of conscious perception (Frith & Frith, 2006). During social interaction, we mostly rely on automatic inferences, whereby social signals are not unbiasedly processed but instead are processed through the lens of pre-acquired priors or implicit assumptions that can be understood as a “practical know-how” for social interaction (Schilbach, 2016). Therefore, social inference can be understood as a special case of unconscious inference. However, social learning has mostly been studied using experimental paradigms of explicit mental state attribution. In fact, some research suggests that mentalization impairments in psychiatric disorder pertain more to automatic and implicit rather than explicit processes (Kronbichler et al., 2019; Langdon, Seymour, Williams, & Ward, 2017; Senju, Southgate, White, & Frith, 2009).

1.4 Neural correlates of social and non-social learning

Social learning refers to the process by which we learn about other people's actions and their consequences (Joiner, Piva, Turrin, & Chang, 2017). In addition to learning the association between others' actions and consequences, we infer on hidden traits (Hackel, Doll, & Amodio, 2015) or intentions (Diaconescu et al., 2017) of others, which engages mentalization. Therefore, an ongoing matter of debate concerns the question of whether learning from others engages the same neural mechanisms as those involved in learning from own experiences such as primary rewards. Alternatively, the account of domain-specificity suggests that social inference relies on distinct brain regions that are specifically evolved for social cognition (Joiner, Piva, Turrin, & Chang, 2017; Lockwood & Klein-Flügge, 2019; Ruff & Fehr, 2014; Wittmann, Lockwood, & Rushworth, 2018).

Model-based approaches to fMRI have been used to elucidate the question of domain-specificity by probing the neural correlates pertaining to learning and decision-making for social and non-social aspects at the same time (cf. section 1.7, Behrens, Hunt, & Rushworth, 2009). Whereas traditional fMRI designs contrast blood oxygen level dependent (BOLD) activity between different conditions of stimulus input or behaviour, computational modelling allows more mechanistic insights by uncovering hidden variables that may underlie the observed data, such as the predictions during decisions and the errors associated with them during outcome processing. In model-based fMRI, parameters that fluctuate on a trial-by-trial basis can be used as predictors of the BOLD response. This way, one can investigate which computations are instantiated by what brain areas at a certain point in time (Behrens, Hunt, & Rushworth, 2009). Indeed, mounting evidence indicates that the neural computations underlying social and non-social learning share a substantial degree of similarity, in addition to being partially dissociable.

1.4.1 Non-social learning

Research investigating non-social learning by means of reinforcement learning models have convergently demonstrated the role of dopamine in reward learning (Daw & Doya, 2006;

Montague, Hyman, & Cohen, 2004; Wolfram Schultz, 2007, 2013). Single-cell recordings in animals and fMRI studies in humans have found reward prediction error signals in dopaminergic neurons of the ventral tegmental area (VTA) and substantia nigra (SN) of the midbrain (Klein-Flügge, Hunt, Bach, Dolan, & Behrens, 2011; Schultz, Dayan, & Montague, 1997; Wolfram Schultz, 2013) and the ventral striatum (vSTR) (Delgado, Nearing, LeDoux, & Phelps, 2008; Garrison, Erdeniz, & Done, 2013; O'Doherty, Cockburn, & Pauli, 2017), which receives dopaminergic projections from the midbrain. Activity in dopaminergic areas increases with positive reward prediction errors (reward larger than predicted) and decreases with negative reward prediction errors (reward smaller than predicted) (Delgado et al., 2008). In addition, the striatum has also been implicated in aversive prediction errors, for instance during pain or fear conditioning (Delgado, Li, Schiller & Phelps, 2008). However, aversive prediction errors mostly activate the amygdala and insula, which is central in risk and error monitoring (Garrison et al., 2013).

Thus, corroborating evidence shows that non-social reward learning is largely dependent on dopamine-signalling in the striatum, which is involved in the evaluation of stimuli and controlling actions pertaining those stimuli in order to make decisions that lead to reward and avoid punishment. Therefore, altered hedonic experience, i.e. anhedonia, which is a core symptom of many psychiatric disorders, is thought to be associated with aberrant dopaminergic signalling (Chekroud, 2015). In addition, striatal activity has been shown to be implicated in the rewarding experiences during social interactions (Pfeiffer et al., 2014). However, the computational and neural underpinnings of social anhedonia and social learning in psychiatric disorders have not yet been investigated.

1.4.2 Social learning

A number of studies have investigated the neural underpinnings associated with social-learning by means of computational modelling (Lockwood & Klein-Flügge, 2019). Social learning is a multifaceted construct, which can either entail learning about the consequences of one's own behaviour on others (e.g. learning about being helpful or liked) or the consequences of other's behaviour on own experiences (e.g. learning about the helpfulness or trustworthiness of someone) (Behrens, Hunt, & Rushworth, 2009; Joiner, Piva, Turrin, & Chang, 2017; Lockwood & Klein-Flügge, 2019; Ruff & Fehr, 2014; Wittmann, Lockwood, & Rushworth, 2018). This project concentrated on studying the computational and neural processes associated in learning to take advice and trust others.

Similarly, previous studies have employed the iterated trust game in order to investigate the neural processes associated with learning to trust others (Delgado, Frank, & Phelps, 2005; Fareri, Chang, & Delgado, 2012; King-Casas et al., 2005; Phan, Sripada, Angstadt, & McCabe, 2010). In the trust game, an investor is given an amount of money of which they can decide to send all, some or nothing to the trustee. The amount sent is usually multiplied before the trustee then decides whether to return all, just a part or none back to the investor. While the investor could earn more by sharing, he/she takes the risk of the trustee not reciprocating the money. Positive social feedback, in this case signalling reciprocated vs. unreciprocated cooperation, recruit the striatum (both the caudate and putamen) as well as the orbitofrontal cortex (OFC) (Delgado, Frank, & Phelps, 2005; Fareri, Chang, & Delgado, 2012; King-Casas et al., 2005; Phan, Sripada, Angstadt, & McCabe, 2010).

King-Casas et al. (2005) elegantly demonstrated that the caudate is involved in building a reputation about the other person during a trust game, as activity in this region that is associated with feedback learning, signals trust during decision phases in later phases of the interaction, after a reputation was built. This is in line with findings of non-social learning showing that the striatum is signalling the probability of primary rewards (Daw & Doya, 2006). Negative violations of social reward, such as unreciprocated *vs.* reciprocated cooperation have been associated with activity in the (anterior) insula (Delgado et al., 2005). This accords with studies demonstrating insula activity in response to misleading advice (Diaconescu et al., 2017) or social exclusion (Eisenberger, Lieberman, & Williams, 2003).

These findings support the notion that social learning rests on domain-general learning structures such as the striatum and insula. However, other studies have pointed to some degree of domain-specificity with regard to social learning: In the seminal work of Behrens, Hunt, Woolrich, & Rushworth (2008), social learning and non-social learning were directly compared. This study used a Bayesian learning model that extended the structure of reinforcement learning models by adopting a dynamic learning rate that scales prediction errors depending on the volatility of the environment, i.e. speed of contingency changes. Participants were asked to learn about the winning probability of two cards. In addition, a social advice was provided at each trial in the form of a red framing of one of the two cards. Participants were previously introduced to the actor that was allegedly giving advice with changing intentions. The model-based analysis revealed that non-social reward prediction errors about the winning card were associated with activity in the vSTR, whereas social prediction errors recruited the dorsomedial prefrontal cortex (dmPFC), middle temporal gyrus (MTG) and temporoparietal junction (TPJ) (Behrens et al., 2008) which are regions that are implicated in ToM and

mentalization (Frith & Frith, 2006). Additionally, the learning rate for both cues correlated with activity in the anterior cingulate, albeit in different sub regions: the reward learning rate was correlated with activity in the sulcus of the anterior cingulate cortex (ACCs) whereas the social learning rate correlated with activity in the gyrus of the anterior cingulate cortex (ACCg).
A modified version of this task was more recently employed in conjunction with a more sophisticated computational approach to learning (Diaconescu et al., 2014, 2017; Sevgi, Diaconescu, Henco, Tittgemeyer, & Schilbach, 2020). Here, Diaconescu et al. (2017) adopted a hierarchical Bayesian model (Hierarchical Gaussian filter) to investigate the neural correlates of hierarchical precision-weighted prediction errors during a task in which participants were asked to infer on an advisors trustworthiness. In this study, precision-weighted prediction errors pertaining the social outcome, were associated with wide spread activation such as the dorsolateral prefrontal cortex (dlPFC), anterior cingulate cortex (ACC), insula and the midbrain (VTA/SN) but also in TPJ and dmPFC which concurs with the results shown by Behrens et al., (2008). Moreover, prediction errors about the volatility of the social advice was represented in the basal forebrain, containing cholinergic neurons. Crucially, this same pattern was found in an earlier task using a perceptual inference task (Iglesias et al., 2013). These findings suggest that apart from domain-general learning computations in the reward circuitry, social learning rests on additional, specified brain regions that may be uniquely *social.*

1.5 Learning and decision-making in psychiatric disorders

The application of reinforcement learning models and Bayesian learning models are central in computational psychiatry because of the main assumption that *disorders of the mind* can be reconstrued as *disorders of learning and decision-making* (Mathys, 2016). Until now, studies on learning and decision-making in psychiatric disorders have largely concentrated on non-social reward learning. For instance, impaired hedonic experience in psychopathology may be captured by altered learning about monetary rewards, which is associated with aberrant striatal signalling (Chekroud, 2015). Despite aberrant social cognition being a central diagnostic criterium for many psychiatric disorders (Schilbach, 2016), the underlying computational mechanisms and how they relate to social anhedonia remain largely unknown. To investigate this, the present PhD thesis investigated learning and decision-making in patients with major depressive disorder (MDD), schizophrenia (SCZ) and borderline personality disorder (BPD). Despite heterogenous presentations of these disorders, they commonly have been associated with impaired learning and decision-making in social and non-social contexts, as outlined in the following.

1.5.1 Major depressive disorder

MDD is characterised by a marked inability to experience reward (anhedonia) and loss of interest in pleasurable activities as well as negative beliefs about the self and others (Weightman, Air, & Baune, 2014). Studies using classical reinforcement learning tasks showed that reduced reward sensitivity in MDD (Harlé, Guo, Zhang, Paulus, & Yu, 2017; Henriques & Davidson, 2000) was associated with blunted activity patterns in the vSTR, which were correlated with symptoms of anhedonia (Gradin et al., 2015; Gradin et al., 2011; Harlé et al., 2017; Rothkirch, Tonn, Köhler, & Sterzer, 2017). Anhedonia is also associated with reduced motivation to engage in social interactions and impaired social cognition (Kupferberg, Bicks, & Hasler, 2016)**.** However, little is known about the computational aspects of social learning in MDD.

Previously, Safra, Chevallier, & Palminteri (2019) found that the severity of depressive symptoms were associated with reduced performance in a learning task when the choices of a co-player were presented, suggesting a negative-audience effect. Studies employing economic games such as the trust game showed that patients with MDD exhibit stronger emotional reactions to unpleasant social interactions and weaker reactions to pleasant social interactions (Robson et al., 2020), which is in line with a negativity bias. Most importantly, patients with MDD show a reduced integration of feedback in social games to adapt behaviour accordingly (Robson et al., 2020). Reduced reward processing has been associated with the notion of learned helplessness or negativity biases (Chekroud, 2015), which may imply overly precise priors of negative predictions (Clark, Watson, & Friston, 2018). Whether these findings pertain to more general reward processing impairments is yet to be investigated.

1.5.2 Schizophrenia

SCZ is a disorder characterised by hallucinations and persecutory delusions (positive symptoms) as well as deficits in emotional expression (negative symptoms). In addition, people suffering from schizophrenia show marked impairments in mentalizing functions (Green, Horan, & Lee, 2015). Studies employing reinforcement learning models provided evidence for impaired reward learning in medicated (Strauss, Waltz, & Gold, 2014; Waltz et al., 2009) and unmedicated patients with SCZ (Juckel et al., 2006; Schlagenhauf et al., 2014), which was associated with blunted responses in the vSTR in response to positive prediction errors, which is in line with findings of patients with MDD. In addition, social interaction tasks such as the trust game have found smaller striatal activity during cooperative responses, suggesting that patients with SCZ may experience social interactions as less rewarding (Gromann et al., 2013; Robson et al., 2020).

While SCZ patients appear to have difficulties in learning about rewarding events, a number of findings point to increased, but not adaptive learning, of non-predicting cues. This is in line with the notion of salience over-attribution to neutral or irrelevant stimuli (Kapur, 2003; Winton-Brown, Fusar-Poli, Ungless, & Howes, 2014). In addition, studies employing the reversal learning task, have demonstrated impairments in tracking the changing probabilistic reward associations indicated by high choice switching behaviour (Culbreth, Gold, Cools, & Barch, 2016; Schlagenhauf et al., 2014). Similarly, studies adopting trust games have shown less strategic decisions, accepting unfair offers less and rejecting fair offers more (Robson et al., 2020).

Thus, not knowing what information to regard and disregard as well as high choice switching may be related to aberrant beliefs pertaining the volatility of the environment (Deserno et al., 2020; Diaconescu et al., 2019; Sterzer, Voss, Schlagenhauf, & Heinz, 2018), which is a more abstract feature of the environment and can be modelled using hierarchical Bayesian models such as the hierarchical Gaussian filter (HGF; (Lomakina et al., 2014; Mathys et al., 2011). In fact, Deserno et al. (2020) showed that patients with SCZ showed increased initial beliefs about volatility that were associated with stronger belief updates. However, with regard to social learning, Diaconescu et al. (2019) suggested that in patients with psychosis, enhanced belief precision is associated with an increased belief rigidity and a decreased propensity to update the model about a confederate. This in contrast would be associated with a reduced estimation of volatility. Although these two suggestions imply different roles for volatility with regard to the social and non-social domains, no study has systematically tested this in patients with SCZ.

1.5.3 Borderline personality disorder

BPD is a complex psychiatric disease that is marked by interpersonal instability as well as emotional and behavioural dysregulation and impaired decision-making (Gunderson, Herpertz, Skodol, Torgersen, & Zanarini, 2018). Patients with BPD show a substantial symptomatic overlap with patients with SCZ, displaying increased self-referential thinking and paranoid ideation. In addition, patients also show increased levels of depression. However, learning and decision-making is less well understood in BPD compared to MDD or SCZ.

Unlike patients with SCZ, patients with BPD showed intact *reversal* learning in non-social environments (Berlin, Rolls, & Iversen, 2005; Dixon-Gordon, Tull, Hackel, & Gratz, 2017), but did show impairments when the context was emotional (Dixon-Gordon et al., 2017). Studies employing the trust game showed that BPD patients' interactions were characterized by reduced trust and less cooperation impeding the maintenance of reciprocal relationship within the experimental setting (King-Casas et al., 2005; Saunders, Goodwin, & Rogers, 2015; Unoka, Seres, Áspán, Bódi, & Kéri, 2009). Interestingly and somehow contradictive, when performing mentalizing tasks, patients with BPD have demonstrated comparable or even improved performance compared to healthy controls (HC) (Fertuck et al., 2009; Frick et al., 2012) and higher confidence in their decisions (Schilling et al., 2012). This may reflect specifically rigid beliefs (overly precise priors) about others (Sharp, 2014), similar to those observed in psychosis. However, patients with BPD often show unstable beliefs about others that polarize between idealization and approach and devaluation and rejection. In a predictive-coding framework this would point to *extreme* beliefs with decreased precision causing an overweighting of prediction errors and constantly changing models of others. Remarkably however, a recent study employing a probabilistic learning task in conjunction with computational modelling found that patients with BPD showed *reduced* learning for social but also non-social cues when they became less predictive of the outcome, i.e. volatile (Fineberg et al., 2018). The authors suggested that this might be due to higher expected baseline volatility in participants with BPD. However, the computational model employed in that study did not explicitly model beliefs about volatility. Moreover, this finding supports the notion of learning impairments in social and non-social contexts, challenging the domain specificity hypothesis.

1.6 Aim of Thesis

The aim of the present PhD thesis was to investigate the computational mechanisms that pertain to probabilistic reward and social learning in healthy controls (HC) and participants with different psychiatric disorders that are known to exhibit social learning and decision-making dysfunction, in particular SCZ, BPD & MDD.

Previous studies probing the neural correlates of learning by means of computational modelling (Behrens et al., 2008, 2007; Diaconescu et al., 2014, 2017) have used paradigms of explicit mental state attribution. However, some research suggests mentalization impairments in psychiatric disorders pertain more automatic rather than explicit processes (cf. section 1.4; Kronbichler et al., 2019; Langdon et al., 2017; Senju et al., 2009). Therefore, the current thesis adopted a probabilistic reward learning task (Sevgi et al., 2020) containing a social cue about which no explicit instruction was given in order to assess the spontaneous use of social information during the learning and decision process. In the task, cards with varying winning probabilities had to be chosen. In addition, the task included a computer-generated face that gazed towards one of these cards providing helpful or misleading advice. In order to directly compare social and non-social inference, we applied parallel reinforcement learning models and hierarchical Bayesian models to behavioural data to obtain a profile of each participant's particular way of updating beliefs about the two types of information. In addition, our modelling framework was specifically designed to quantify the relative weighting of predictions about social and non-social information.

1.6.1 Neural correlates of social and non-social learning and decision-making

In the fMRI study, I used the learning trajectories as well as the weighting factor from the best performing model, the HGF (Mathys et al., 2014; Mathys et al., 2011) and used them as predictors in a model-based fMRI analysis. This approach was employed to investigate the neural correlates of social and non-social inference in healthy participants. More specifically, we asked whether social learning signals (i.e. prediction errors) during uninstructed inference would yield neural activations similar to those found in studies of instructed inference (Behrens et al., 2008; Diaconescu et al., 2017). In addition, we probed the neural correlates of social and non-social predictions during choice and evaluated whether inter-individual variance in the propensity to use the social cue during decision-making would be reflected in differential neural activity.

1.6.2 Transdiagnostic mechanisms of social and non-social learning and decision-making

As outlined in chapter 1.5., Bayesian accounts of psychiatric disorders argue for an imbalance between belief precision and data precision that give rise to aberrant inference and maladaptive beliefs (Haker et al., 2016; Mathys, 2016; Stephan & Mathys, 2014). During the past years, these accounts have started to be formally tested (e.g. Deserno et al., 2020; Lawson, Rees, & Friston, 2014), albeit with a focus on non-social inference. Given the ubiquity of social impairments in psychiatric disease, the current study investigated learning and decision-making in a social context. More specifically, since our paradigm consisted of social and non-social cues, we aimed to systematically investigate whether potential aberrances in learning pertain certain aspects of the environment or rather general process that are independent of the domain. To test this, we used the learning trajectories of the winning model (HGF) to extract precision weights (i.e. dynamic learning rates) when learning about social and non-social contingencies and their volatility to investigate inference style in both domains at the same time. In addition, we investigated whether patient groups differed in the extent to which social predictions are weighted during decisions and whether this is reflected in scores of social anhedonia.

2 Bayesian modeling captures inter-individual differences in social belief computations in the putamen and insula

This chapter includes the first study that used model-based fMRI in healthy controls to investigate the neural activity associated with social and non-social predictions and inter-individual differences in the propensity to weight social over non-social predictions. The findings highlight the role of the insula in tracking both social and non-social predictions during decision-making and in signalling prediction errors during learning. Moreover, the results showed that individual differences in the extent to which participants weighted their social predictions were correlated with activity in the putamen and insula. These findings demonstrate the usefulness of model-based fMRI in uncovering the behavioural and neural mechanisms of spontaneous social cue integration in learning and decision-making. The manuscript was accepted in Cortex in 2020.

Authors:
Henco L, Brandi M-L, Lahnakoski JM, Diaconescu AO, Mathys C & Schilbach L.

Contributions:
The author of this thesis is the first author of the publication; AOD and LS designed the research; **LH** and MLB implemented the experiment with the MR environment and collected the data; **LH** performed the computational data analysis with help of AOD, CM and the fMRI analysis with help of MLB and JML; **LH** wrote the manuscript; all authors reviewed and edited the manuscript. LS was the supervisor of the project and provided funding.

Bayesian modelling captures inter-individual differences in social belief computations in the putamen and insula

Lara Henco[a,b], Marie-Luise Brandi[a], Juha M. Lahnakoski[a], Andreea O. Diaconescu[c,d,e], Christoph Mathys[d,f,g*] & Leonhard Schilbach[a,b,h,i*]

[a] Independent Max Planck Research Group for Social Neuroscience, Max Planck Institute of Psychiatry, Munich, Germany
[b] Graduate School for Systemic Neurosciences, Munich, Germany
[c] Department of Psychiatry (UPK), University of Basel, Basel, Switzerland
[d] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland
[e] Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), University of Toronto, Canada
[f] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
[g] Interacting Minds Centre, Aarhus University, Aahrus, Denmark
[h] Department of Psychiatry, Ludwig-Maximilians-Universität, Munich, Germany
[i] Outpatient Clinic and Day Clinic for Disorders of Social Interaction, Max Planck Institute of Psychiatry, Munich, Germany

Corresponding Author: Lara Henco, Independent Max Planck Research Group for Social Neuroscience, Max Planck Institute of Psychiatry, Kraepelinstr. 8-10, 80804 Munich, Germany, lara_henco@psych.mpg.de

* The authors contributed equally to this work

Abstract

Computational models of social learning and decision-making provide mechanistic tools to investigate the neural mechanisms that are involved in understanding other people. While most studies employ explicit instructions to learn from social cues, everyday life is characterized by the spontaneous use of such signals (e.g. the gaze of others) to infer on internal states such as intentions. To investigate the neural mechanisms of the impact of gaze cues on learning and decision-making, we acquired behavioural and fMRI data from 50 participants performing a probabilistic task, in which cards with varying winning probabilities had to be chosen. In addition, the task included a computer-generated face that gazed towards one of these cards providing implicit advice. Participants' individual belief trajectories were inferred using a hierarchical Gaussian filter (HGF) and used as predictors in a linear model of neuronal activation. During learning, social prediction errors were correlated with activity in inferior frontal gyrus and insula. During decision-making, the belief about the accuracy of the social cue was correlated with activity in inferior temporal gyrus, putamen and pallidum while the putamen and insula showed activity as a function of individual differences in weighting the social cue during decision-making. Our findings demonstrate that model-based fMRI can give insight into the behavioural and neural aspects of spontaneous social cue integration in learning and decision-making. They provide evidence for a mechanistic involvement of specific components of the basal ganglia in subserving these processes.

Keywords: Learning and decision-making, Social inference, Bayesian modelling, fMRI

1. Introduction

Successful social interaction requires learning from others and making decisions that in turn lead to rewarding experiences. Although similar to reward learning in non-social contexts, social learning is thought to engage different processes by which not only reward associations are learned, but also the hidden traits (Hackel, Doll, & Amodio, 2015) or states (e.g. intentions) (Diaconescu et al., 2017) which may modulate these associations. Accordingly, social learning has been found to engage brain regions that may have a unique role in social cognition in addition to the neural circuitry involved in non-social learning (Joiner, Piva, Turrin, & Chang, 2017; Lockwood & Klein-Flügge, 2019; Ruff & Fehr, 2014; Wittmann, Lockwood, & Rushworth, 2018).

Reinforcement learning studies have repeatedly found that striatal activity is associated with non-social reward prediction errors, i.e. the difference between actual and expected reward (cf. Dayan & Daw, 2008; O'Doherty, Cockburn, & Pauli, 2017), but also reward prediction errors in various social contexts (e.g. Báez-Mendoza & Schultz, 2013; Burke, Tobler, Baddeley, & Schultz, 2010; Hackel et al., 2015; Lockwood, Apps, Valton, Viding, & Roiser, 2016; Lockwood & Klein-Flügge, 2019). For instance, in trust games in which participants are required to make risky investments with other players, parts of the striatum including the caudate and putamen show stronger activations in response to reciprocated cooperation (Delgado, Frank, & Phelps, 2005; Fareri, Chang, & Delgado, 2012; King-Casas et al., 2005). Activity in these regions is also associated with reward predictions about others during trust decisions (King-Casas et al., 2005; Diaconescu et al., 2017). Negative violations of social reward, such as unreciprocated cooperation (Rilling, King-Casas, & Sanfey, 2008), misleading advice (Diaconescu et al., 2017) and social exclusion (Eisenberger, Lieberman, & Williams, 2003) have been associated with activity in the insula, which is also involved in risk and error monitoring in non-social contexts (cf. Iglesias et al., 2013).

In addition, some brain regions may be more strongly involved in social learning than in non-social learning. For instance, paradigms in which participants were asked to learn about the trustworthiness of a partner through trial and error (Behrens, Hunt, Woolrich, & Rushworth, 2008; Diaconescu et al., 2017; King-Casas et al., 2005) have been used to show that social prediction errors engage brain areas previously associated with mentalization, such as the temporoparietal junction (TPJ) and the dorsomedial prefrontal cortex (dmPFC). Other studies highlighted the domain specificity of the anterior cingulate gyrus (ACCg) when learning from others (Apps, Lesage, & Ramnani, 2015; Apps, Rushworth, & Chang, 2016; Lockwood, Apps, Roiser, & Viding, 2015).

The majority of studies which investigated the neural correlates of learning the trustworthiness of others, thereby probing mentalization, instructed participants explicitly to learn from a partner's advice (Behrens et al., 2008; Diaconescu et al., 2017). Most everyday life social interactions, however, require us to automatically infer on mental states by using nonverbal signals such as gaze behaviour (Schilbach et al., 2013). Therefore, in the current study, we decided to investigate the neural mechanisms of uninstructed social learning and decision-making by means of functional magnetic resonance imaging (fMRI).

To this end, we employed an established probabilistic learning task (Sevgi, Diaconescu, Henco, Tittgemeyer, & Schilbach, 2020) in which participants can learn from two types of information, i.e. a non-social cue (cards with different colours) and a social cue (gaze shift of a face presented in the centre of the screen), in order to maximize the reward associated with a card draw (Figure 1A). In this task, participants were not explicitly instructed to pay attention to the face in order to probe the spontaneous use of social information. Three types of computational models of learning and decision-making were used to fit participants' choices. These models varied in their complexity of the belief updating process and have been employed in previous studies of learning under uncertainty (DeBerker et al., 2016; Iglesias et al., 2013). Furthermore, the modelling framework was constructed in such a way that it allowed us to estimate the relative weight participants were affording to their learned beliefs about the social cue compared to the non-social cue when predicting the outcome of the task. We also captured the usage of the social cue by means of model-agnostic measures, i.e. subjective post-experimental reports as well as gaze fixations during decision-making by means of simultaneous eye-tracking.

The learning trajectories as well as the weighting factor from the best performing model, the hierarchical Gaussian filter (HGF; Mathys et al., 2014; Mathys, Daunizeau, Friston, & Stephan, 2011), were used as predictors in model-based fMRI analysis to uncover the neural mechanisms of social and non-social learning and decision-making. We evaluated whether social learning signals during uninstructed inference would yield neural activations similar to those found in studies of instructed inference (Behrens et al., 2008; Diaconescu et al., 2017). This allowed us to evaluate whether inter-individual variation in the propensity to use the social cue during decision-making is reflected in differences of neural activity. We expected the striatum to be involved in the representation of social cue probabilities and were specifically interested in investigating whether individual differences in weighting the social over non-social information in the task were also represented in this part of the brain. We further evaluated the estimated uncertainty for social and non-social cues during decision-making. We predicted that the insula would code both social and non-social uncertainty and asked whether social uncertainty is additionally tracked by regions involved in mentalization. Furthermore, we probed the neural correlates of social and non-social prediction errors and predicted to find overlapping activations in the anterior cingulate and insula as well as activations associated with social learning in brain regions involved in mentalizing, such as the TPJ and the dmPFC (Behrens et al., 2008; Diaconescu et al., 2017).

2. Methods

2.1 Participants

A total of 55 healthy volunteers (28 female; mean age 25.2± 5.6 years, range: 18 – 48 years) participated in the study. These participants were recruited through the Max Planck Institute of Psychiatry as well as local universities. They were all right-handed, had normal or corrected-to-normal vision and reported no history of neurological or psychiatric disease. Furthermore, they did not meet any contraindications for magnetic resonance imaging (MRI) measurement, such as metal implants or claustrophobia. All participants stated to be non-smokers and none of them reported current intake of psychoactive medication. All participants were naïve to the purpose of the experiment and provided informed consent to take part in the study after a written/verbal explanation of the study procedure. Participants received a reimbursement for participation and an additional amount of money (1-6 Euro) that depended on their score in the task. The study was in line with the Declaration of Helsinki and approval for the experimental protocol was granted by the local ethics committee of the Medical Faculty of the Ludwig-Maximilians-University of Munich. Five measured participants were not included in the analysis: two were excluded due to abnormalities in the structural brain scans, one due to technical issues with the task presentation on the scanner monitor, one participant did not perform the task according to the instruction, and one participant was excluded because an exclusion criterion (nicotine abuse) applied, which was communicated subsequent to measurement. Accordingly, we analysed data from 50 participants (25 female; mean age 24.8 ± 5 years, range: 18 – 48 years).

2.2 Experimental paradigm and procedure

Participants completed a probabilistic learning task, comprising a non-social and a social cue (Figure 1A). The task, initially introduced by (Sevgi et al., 2020), consisted of 120 trials and lasted approximately 20 min. Participants were instructed to choose one of two cards (green or blue) on every trial and were told that the winning probability of the colours would change throughout the task. A computer-generated face was presented at the centre of the screen during the entire trial. At the trial start, the face looked down, then raised its eyes to look directly at the participant, and then shifted its gaze towards one of two cards presented on either side of it (Figure 1A). Independently of the winning probability of the card colours, the probability of the face gazing towards the winning card, thus providing a social cue, was also systematically manipulated. Participant choice was enabled two seconds after the gaze shift of the face and lasted until a response was made. Trials were not counted if the participant pressed a button before the choice was enabled or if they took more than 5 seconds to respond after the choice was activated. In these cases, the screen showed "response too early/late" and the outcome of the choice was not displayed. The choice phase was followed by a jittered delay (2–4 seconds) before the outcome (correct/wrong) was presented for 2 seconds. During choice, both cards were showing reward values (ranging from 1-9), which were added to a cumulative score that was presented during the feedback phase if the participant chose the correct card. When the answer was wrong, the score remained the same. Participants were told that the numbers were sampled randomly and that they were not associated with the winning probabilities of the cards. Participants were told that if they were completely uncertain about the winning probabilities, they might want to pick the card associated with a higher reward value. The outcome was signalled to the participant by a green check mark (correct choice) or a red cross (incorrect choice. All trials were separated by a jittered inter-trial interval (3-6 seconds) and 12 of these inter-trial intervals were jittered at longer durations (12-15 seconds), similar to including null trials.

Prior to the task, participants were informed that the card winning probabilities would change during the task. Participants were not explicitly instructed to learn about the social cue, but were merely told that the face in the centre of the screen was included to make the task more interesting. The probability schedule of the social cue was orthogonal to the non-social cue as shown in Figure 1B. During the first half of the experiment, the winning probability of the blue card was stable at 75% (trials 1–60), followed by a volatile period where winning probability changed from 20% (trials 61-80; 101-120) to 80% (trials 81-100). The gaze schedule started with a stable phase with 75% accuracy (trials 1–40), followed by a volatile period where gaze accuracy changed from 20% (trials 41-50; 61-70) and 80% (trials 51-60; 71-80). During trials 80-120 the gaze accuracy had a probability of 12%. For 8 participants, who were recruited during the pilot phase of the study, the volatile phase of the social cue started 10 trials later. The paradigm was presented by Presentation software (Presentation Version 16.3, Build 12.20.12, Neurobehavioural Systems Inc., Berkeley, California, USA, www.neurobs.com) running on a Microsoft Window XP operating system and stimuli were presented on a 30-inch LCD OptoStim H-3/30 Medres MRI compatible monitor on a background of grey luminance with a resolution of 1024x768 and a refresh rate of 60 Hz. Participants responded to Stimuli using two buttons on a response box (LSC-400B controller, Lumina, Cedrus).

Prior to the MRI session, participants were asked to answer a standard set of questionnaires used in the research group. It included the autism quotient (AQ; Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) emotional quotient (EQ; Anticipatory and Consummatory Interpersonal Pleasure Scale (ACIPS; Gooding & Pflum, 2014), Liebowitz Social Anxiety Scale (LSAS; Liebowitz, 1987), the Becks Depression Inventory (BDI-II; Kühner, Bürger, Keller, & Hautzinger, 2006), the Social Network Questionnaire (SNQ; Linden, Lischka, Popien, & Golombek, 2007), the Toronto Alexythimia Scale (TAS; Bagby, Taylor, & Parker, 1994) as well as the Reading the Mind in the Eyes Test (RMET; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). The psychometric data was analysed within the scope of a different study. In addition, participants filled out a post-experimental questionnaire to assess the subjective learning experience during the task, asking how difficult the task was (from 0 to 100), how much they used the gaze (from 0 to 100) and how much it helped them during the task (from 0 to 100). The results of the post-experimental questionnaire can be seen in the appendix (Table A. 1).

2.3 Computational Modelling

The modelling approach followed the "observing the observer" framework in which two types of models (perceptual and response models) are paired in order to allow the inference of an observer (i.e., the experimenter) on the inference of a participant: Perceptual models describe the participant's belief trajectories about the hidden causes (states) of the sensory inputs (*here:* social and non-social cue); the response models describe how these beliefs are translated into decisions (Daunizeau et al., 2010).

2.3.1 Perceptual Models

We used 3 perceptual models that had been employed in previous studies (cf. Iglesias et al., 2013) and that varied with regard to the complexity of the belief updating process. Perceptual model 1 comprises two parallel hierarchical Gaussian filters (HGF; Mathys et al., 2014; Mathys et al., 2011), which are inversions of generative models of the sensory inputs the participant experiences, i.e. card and gaze outcomes (Figure 2). This approach assumes that participants are dynamically updating their beliefs (i.e., posterior probability distributions) in order to infer on the hidden environmental states x that cause the experienced sensory inputs. In the generative model, these “to-be-inferred-on” states are coupled in a three-level hierarchy: The lowest level $x_{1_{gaze}}$ represents the accuracy of the gaze in a binary form (1=correct, 0=incorrect), level $x_{2_{gaze}}$ represents the tendency of the gaze to be correct or incorrect and level $x_{3_{gaze}}$ represents the volatility of this tendency to be accurate. Correspondingly, the lowest level $x_{1_{card}}$ represents the accuracy of the blue card in a binary form (1=correct, 0=incorrect), level $x_{2_{card}}$ represents the tendency of the blue card to be correct or incorrect and level $x_{3_{card}}$ represents the volatility of the tendency of the blue card to be correct. The third state evolves as a first-order autoregressive (AR(1)) process. The second state evolves as a Gaussian random walk with a step size determined by the state at the third level. The probability of x_1 is a sigmoid transformation of x_2.

$$p(x_1 = 1) = \frac{1}{1+\exp(-x_2)} \qquad (1)$$

Given trial-wise responses of participants that indicated whether they had followed the advice implicit in the gaze, this model was inverted in order to infer participant-specific parameters and belief trajectories (Mathys et al., 2014). This resulted in belief trajectories at three hierarchical levels $i = 1,2,3$. The beliefs $\mu_i^{(k)}$ about the state of the environment are updated on every trial k via prediction errors $\delta_{i-1}^{(k)}$ from the level below weighted by a precision ratio (Equations 2-3) where the beliefs’ precision $\pi_i^{(k)}$ on each level is equal to the inverse variance of the belief $\pi_i^{(k)} = 1/\sigma_i^{(k)}$. Thus, the precision ratio causes larger belief updates when the precision of the posterior belief is low and the precision of the data is high.

$$\Delta\mu_i^{(k)} \propto \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \delta_{i-1}^{(k)} \qquad (2)$$

$$\Delta\mu_2^{(k)} \propto \frac{1}{\pi_2{}^{(k)}} \delta_1^{(k)} \qquad (3)$$

$$\Delta\mu_3^{(k)} \propto \frac{\hat{\pi}_2^{(k)}}{\pi_3{}^{(k)}} \delta_2^{(k)} \qquad (4)$$

The evolution of beliefs is governed by participant-specific parameters: ω_{2card} and ω_{2gaze} determine the participant-specific evolution rate at the second level. As such, they describe how fast contingencies of gaze and card stimuli with outcome change in general, independent of phasic spikes and dips. ω_{3card} and ω_{3gaze} play the corresponding role at the third level, representing the evolution rates of the volatilities of the contingencies. Refer to table A2 in the appendix for configurations of priors used in parameter estimation.

Perceptual model 2 is a parallel (gaze and card) version of the Sutton K1 model which assumes a learning rate that varies over time as a function of recent prediction errors (Sutton, 1992). Perceptual model 3 is a parallel classical reinforcement learning model which assumes a learning rate that is fixed and participant-specific (Rescorla & Wagner, 1972).

2.3.2 Response Models

In all response models, a combination of first level predictive beliefs about gaze $\hat{\mu}_{1,gaze}^{(t)}$ and card $\hat{\mu}_{1,card}^{(t)}$contingency with outcome (called ‘accuracy’ in what follows), weighted by precision was mapped onto decisions (Equation 4). The combined belief was modelled as the sum of the posterior predictive expectation of gaze accuracy $\hat{\mu}_{1,gaze}^{(t)}$ and card accuracy $\hat{\mu}_{1,card}^{(t)}$ weighted by weights $w_{gaze}^{(t)}$ and $w_{card}^{(t)}$ (Equation 5 & 6), which are a function of the precisions of gaze and card accuracy predictions, respectively. Since beliefs were modelled in the gaze space (i.e., all cues and outcomes were parameterized with respect to the card receiving the gaze), the posterior predictive expectation of card $\hat{\mu}_{1,card}^{(t)}$ was translated into gaze space, so that $\hat{\mu}_{1,card}^{(t)} = \hat{\mu}_{1,card}^{(t)}$ if the gaze went to the blue card, but $\hat{\mu}_{1,card}^{(t)} = 1 - \hat{\mu}_{1,card}^{(t)}$ if the gaze went to the green card. The precisions $\hat{\pi}_1$ (Equation 7 & 8) were calculated as the inverse variances of a Bernoulli distribution of the posterior card and gaze estimates at the first level of the hierarchy. This entails that precision increases when $\hat{\mu}_1^{(t)}$ moves away from 0.5. The constant parameter $\zeta > 0$ is a weight on the precision of gaze accuracy representing the relative sensitivity of a participant to the social input compared to the non-social input. Simulations reported in Figure 3 illustrate the implications of high and low ζ values for decision-making.

$$b^{(t)} = w_{gaze}^{(t)}\, \hat{\mu}_{1,gaze}^{(t)} + w_{card}^{(t)}\, \hat{\mu}_{1,card}^{(t)} \qquad (4)$$

$$w_{gaze}^{(t)} = \frac{\zeta \hat{\pi}_{1,gaze}^{(t)}}{\zeta \hat{\pi}_{1,gaze}^{(t)} + \hat{\pi}_{1,card}^{(t)}} \qquad (5)$$

$$w_{card}^{(t)} = \frac{\hat{\pi}_{1,card}^{(t)}}{\zeta \hat{\pi}_{1,gaze}^{(t)} + \hat{\pi}_{1,card}^{(t)}} \qquad (6)$$

$$\hat{\pi}_{1,gaze}^{(t)} = \frac{1}{\hat{\mu}_{1,gaze}^{(t)}\,(1 - \hat{\mu}_{1,gaze}^{(t)})} \qquad (7)$$

$$\hat{\pi}_{1,card}^{(t)} = \frac{1}{\hat{\mu}_{1,card}^{(t)}\,(1 - \hat{\mu}_{1,card}^{(t)})} \qquad (8)$$

We coded participants’ responses y in terms of congruency with the ‘advice’, that is, whether participants chose the card that was indicated by the gaze shift (1) or not (0). In the response model, the probability of following the advice $Prob_{gaze}^{(t)}$ was modelled as a logistic sigmoid (softmax) function of combined belief $b^{(t)}$ (Equation 4), weighted by the expected reward of the card when following the advice r_{gaze} or not $r_{notgaze}$ (Equation 9).

$$prob_{gaze} = p\left(y^{(t)} = 1\right) = 1/\left(1 + \exp\left(-\gamma^{(t)}\left(r_{gaze}^{(t)} b^{(t)} - r_{notgaze}^{(t)}\left(1 - b^{(t)}\right)\right)\right)\right) \qquad (9)$$

The extent to which a participant's beliefs map onto actions is dependent on inverse decision temperature $\gamma^{(t)}$. A larger $\gamma^{(t)}$implies a more deterministic relationship between actions and belief whereas a smaller $\gamma^{(t)}$ is indicative of a weaker relationship and more erratic or stochastic behaviour. We implemented four different versions of $\gamma^{(t)}$ to test different hypotheses (mechanisms) of belief-to-response mapping. We inverted models in which $\gamma^{(t)}$ was either (1) a combination of the log-volatility of the third level for both gaze and card combined with constant participant-specific decision noise β (Equation 10), (2) a combination of the log-volatility of the third level for gaze and participant-specific decision noise (Equation 11), (3) a combination of the log-volatility of the third level for card and participant-specific decision noise (Equation 12) or (4) the participant-specific decision noise alone (Equation 13).

1) $\gamma^{(t)} = \beta \exp\left(-\hat{\mu}_{3,card}^{(t)} - \hat{\mu}_{3,gaze}^{(t)}\right)$ (10)

2) $\gamma^{(t)} = \beta \exp\left(-\hat{\mu}_{3,gaze}^{(t)}\right)$ (11)

3) $\gamma^{(t)} = \beta \exp\left(-\hat{\mu}_{3,card}^{(t)}\right)$ (12)

4)$\gamma^{(t)} = \beta$ (13)

2.3.3 Combination of perceptual and response models

Overall, we used six different models to model learning and decision-making: the HGF was combined with all four response models. Due to the lack of a third level, the Sutton K1 and Rescorla Wagner models were only combined with the response model 4 in which decision noise was a participant-specific decision noise parameter (Equation 13). We used the HGF toolbox version 4.1, which is part of the software package TAPAS (https://translationalneuromodeling.github.io/tapas). A quasi-Newton optimization algorithm was employed for estimation.

2.4 Model selection

For model comparison we used the log model evidence (LME), which is calculated in the HGF Toolbox during estimation and represents a trade-off between model complexity and model fit. The LME values for each of the 6 model configurations for each participant were subjected to random-effects Bayesian Model Selection (spm_BMS in SPM12 ; www.fil.ion.ucl.uk/spm) to find the expected posterior probabilities (EXP_P), i.e. the probability for each model of it having generated the responses for a randomly chosen participant out of all models in the model space. We also report the exceedance probability (XP) and protected exceedance probability (PXP), i.e. the probability that a given model better explains the data than any other model in the comparison space (Stephan et al., 2009; Rigoux et al., 2014).

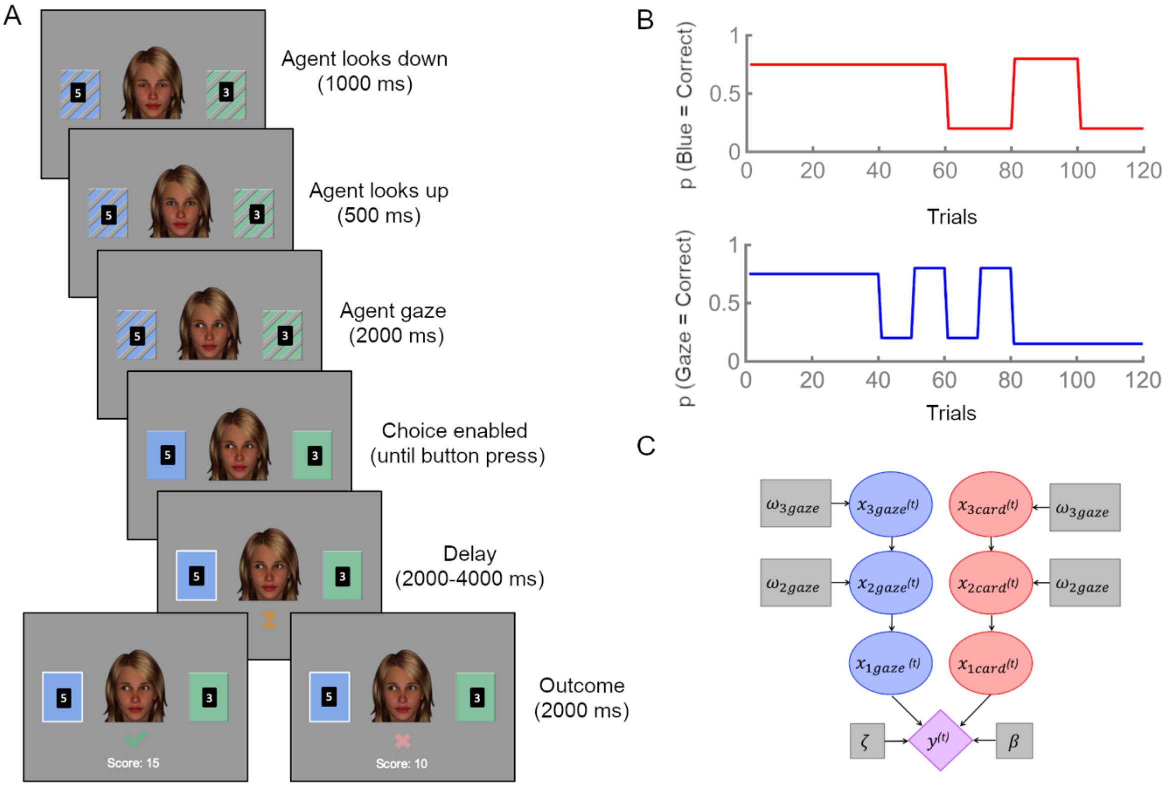


Figure 1A: Trial flow and task design. On every trial, participants choose one of two cards (green & blue). After the choice is logged, an hourglass is presented followed by a green tick or a red cross depending on whether the response was correct or wrong. With every correct response, the score of the chosen card is added onto a cumulative score that participants were instructed to maximize and which determined the additional amount (1–6 euro) paid to the participant at the end of the experiment. **B:** Probability schedule of the social (blue) and non-social (red) cue. **C:** Two parallel learning systems that describe participants' learning about the probability and volatility of the social (blue) and non-social (red) cues. The circles (blue and red) and the diamond (purple) represent states that change in time (i.e. trial t), whereas the squares denote parameters estimated across time (see Methods 2.3).

2.5 Behavioural Analysis

As a proof-of-concept analysis for our computational parameter ζ (i.e. the weighting of gaze input), we correlated this parameter with subjective reports given in response to a post-experimental questionnaire, asking participants how much they used the gaze (on a scale from from 0 to 100) and how much it had helped them during the task (on a scale from 0 to 100). Also, we tested the association with the parameter ζ and the percentage of trials in which a participant chose the card that had been indicated by the gaze. In addition, advice taking behaviour (card chosen indicated by gaze) was subjected to a repeated measures ANOVA with Task Phase as within-subject factor (gaze accuracy high vs. gaze accuracy volatile vs. gaze accuracy low) and ζ as covariate. Statistical tests were performed using JASP (Version 0.9; https://jasp-stats.org/) and Matlab (Version 2018a; www.mathworks.com).

2.6 fMRI acquisition and preprocessing

fMRI data were acquired with a 3-Tesla MR imaging system (MR750, GE, Milwaukee, USA) using a 32-channel head coil. Anatomical screening was performed acquiring T1-weighted 3D inversion recovery fast spoiled gradient-echo scans with a voxel size of 1x1x1 mm. Whole brain functional images were acquired (AC-PC-orientation, interleaved bottom-up, slice number = 40, inter-slice gap = 0.5 mm, TE = 20 ms, TR = 2000 ms, flip angle=90°, voxel size = 3 × 3 × 3 mm, FOV 24 × 24 cm, matrix 96 × 96, resulting in-plane resolution 4 × 4 mm). Each run lasted approximately 30 min, resulting in around 900 volumes.

Preprocessing of fMRI data was performed using MATLAB and SPM12 (Statistical Parametric Mapping Software, www.fil.ion.ucl.ac.uk/spm). Slice time correction was applied to account for the order of initially acquired interleaved slices. Using rigid body transformation, images were then spatially realigned to the volume mean and 6 motion regressors were obtained, which were later used as nuisance regressors in the GLM. The participant's structural scan was then co-registered to the volume mean. The co-registered structural image was segmented and parameters obtained by this process were applied for normalising functional and structural images to the Montreal Neurological Institute (MNI) standard template with a voxel resolution of 2 x 2 x 2 mm for functional images and 1 x 1 x 1 mm for structural images. In addition, to account for respiratory, cardiac, or vascular activity, a CompCor analysis was performed using the PhysIOtoolbox (Kasper et al., 2017; https://translationalneuromodeling.github.io/tapas). Using this method, time courses of voxels within WM and CSF (masks obtained from segmentation) were extracted from the smoothed images and subjected to a principal components analysis. The first three principal components of both WM and CSF entered the GLM as nuisance regressors as well as six movement parameters generated by the realignment step. For each nuisance regressor, we also included the absolute first order derivate. Due to losing the structural scan of one subject when transferring data (after preprocessing), the GLM of one participant only contained the twelve motion nuisance regressors, without the principal components of WM and CSF.

2.7 fMRI analysis

First-level

In our neuroimaging analysis we investigated the neural correlates of the following computational trajectories: The belief about the probability of the gaze to give correct advice ($\hat{\mu}_{1,gaze}^{(t)}$), the variance (i.e. uncertainty) of this belief ($\widehat{\sigma}_{1,gaze}^{(t)}$), and the variance about the probability of the winning card colour ($\widehat{\sigma}_{1,card}^{(t)}$). We did not use $\hat{\mu}_{1,card}^{(t)}$ in the analysis since we didn't expect neural activity with regard to the winning probability of the blue or green card (the coding of blue = 1 and green = 0 was arbitrary). In addition, we investigated the neural correlates of the social prediction error signal $\delta_{1gaze}^{(t)}$ and the non-social prediction error signal $\delta_{1card}^{(t)}$ (an example of these trajectories can be seen in Figure 2).

In order to investigate whether neural activity change was associated with these parameters, we defined voxel-wise general linear models (GLMs) on the first level of analysis. In the main GLM analysis the choice phase was modelled starting from the time point of the gaze shift until the response of the participant. The choice phase was parametrically modulated with the participant-specific belief trajectories $\hat{\mu}_{1,gaze}^{(t)}$. The outcome phase of the task (modelled for 2 seconds starting at outcome presentation) was parametrically modulated by four regressors: The first regressor contained $\delta_{1gaze}^{(t)}$ neutralised, where the choice was wrong by setting the regressor's value to zero. In the second regressor, $\delta_{1gaze}^{(t)}$ was set to zero where the choice was correct. This way we could evaluate $\delta_{1gaze}^{(t)}$ for wrong, correct, and all choices. This was important since the surprise about the social cue has a different relevance depending on whether the participant's choice was correct or wrong. Therefore, misleading advice that preceded a correct choice might be differently valenced than misleading advice that preceded a wrong choice. According to the same rationale, the third and fourth regressors contained $|\delta_{1card}^{(t)}|$ neutralized where the gaze was correct and where it was incorrect, respectively. The absolute value of prediction error was chosen because it was an arbitrary choice whether to code blue outcomes as 1 and green ones as 0 or the other way around. In this analysis, we also examined the prediction error signal for all trials, irrespective of social cue accuracy and separately for trials in which the social cue was correct or wrong. Due to a correlation between $\hat{\mu}_{1,gaze}^{(t)}$ and $\widehat{\sigma}_{1,gaze}^{(t)}$, we estimated $\widehat{\sigma}_{1,gaze}^{(t)}$ and $\widehat{\sigma}_{1,card}^{(t)}$ in a separate GLM, which was the same as the one described above but differed in that the choice phase was modulated by $\widehat{\sigma}_{1,gaze}^{(t)}$ and $\widehat{\sigma}_{1,card}^{(t)}$ and not by $\hat{\mu}_{1,gaze}^{(t)}$. For completeness, we also estimated a GLM that included all parametric regressors ($\hat{\mu}_{1,gaze}^{(t)}$, $\widehat{\sigma}_{1,gaze}^{(t)}$ and $\widehat{\sigma}_{1,card}^{(t)}$) as modulators of the choice phase (cf. appendix).

To investigate the neural correlates of fixations (see Methods, section 2.8 for acquisition and analysis) on the face during choice, we defined another GLM, in which the choice regressor was parametrically modulated by the fixation proportions on the face area. This GLM was estimated for 44 participants, as some participants had to be discarded due to insufficient quality of the eye tracking data (i.e. blurred corneal reflection).

In all GLMs, we modelled missed responses with separate regressors and all regressors were convolved with a canonical hemodynamic function. In addition, all parametric regressors were z-scored and not orthogonalized.

Second-level

Contrast images for each parametric modulator were estimated at the first level against baseline. These contrast images were entered into a second level one-sample t-test for group level inference and we examined positive and negative effects of the contrasts. We also compared positive ($\delta_{1gaze}^{(t)} > 0$, i.e., gaze helpful) and negative social prediction errors ($\delta_{1gaze}^{(t)} < 0$, i.e., gaze misleading) directly, by entering subject-wise pairs of positive and negative contrast images of the parametric modulator containing the prediction error signal $\delta_{1gaze}^{(t)}$ into a paired t-test. We also directly compared negative social prediction errors during incorrect outcomes (i.e. participant followed misleading gaze) with negative social prediction errors during correct outcomes (i.e. participant didn’t follow misleading gaze) as well as positive social prediction errors during correct outcomes (i.e. participant followed helpful gaze) with positive social prediction errors during incorrect outcomes (i.e. participant didn’t follow helpful gaze).

To examine individual differences in brain areas associated with $\hat{\mu}_{1,gaze}^{(t)}$, we included the social weighting factor ζ estimated from the winning computational model as a variable of interest in the respective t-tests. As a non-computational equivalent, we used the subjective report of the post-experimental questionnaires (Tab. A1), stating the extent to which participants used the gaze during the task. In this analysis, we included 48 participants since the data of two participants was missing (Tab. A1). Since ζ and the post-experimental questionnaire were correlated (Fig. 2a), these two were entered separately in the second level analysis. To examine individual differences in brain regions correlated with -$\widehat{\sigma}_{1,card}^{(t)}$ as a function of weighting the non-social cue, we used $-\zeta$ for the computational covariate and $-Question3$ for the questionnaire covariate. Clusters were formed at uncorrected p = 0.001, followed by a cluster-level correction for multiple testing, with significance defined as cluster-level p-values < 0.05 after correction for family-wise error rate (FWE).

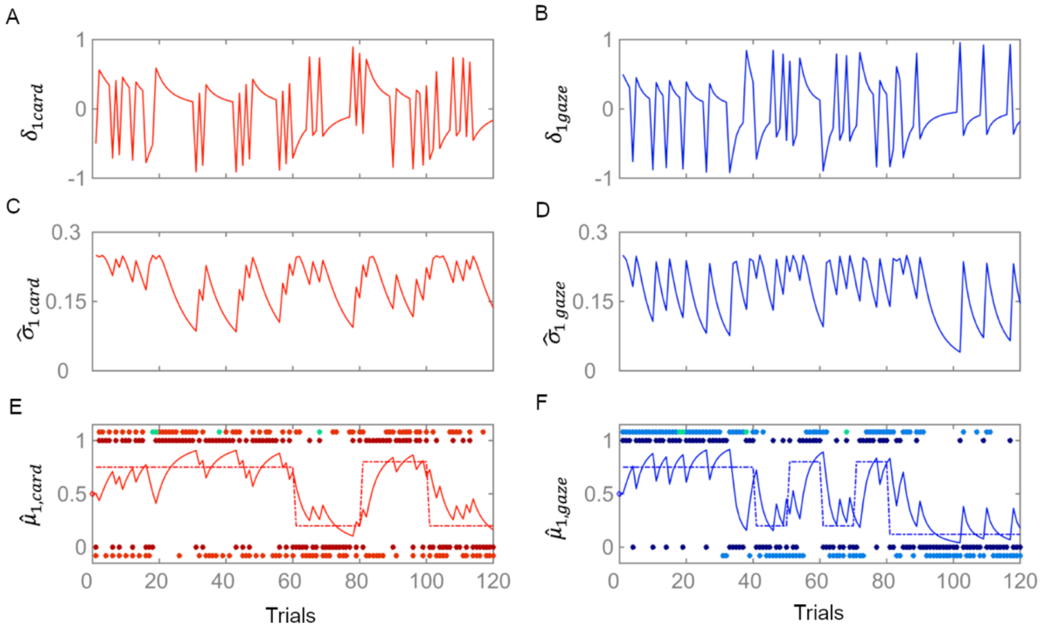


Figure 2. Example of participant-specific learning trajectories for both cues. A) Prediction error δ_{1card} (red) about trial outcome in terms of the non-social cue and B) prediction error δ_{1gaze} (blue) about the trial outcome in terms of the social cue. C) Variance (uncertainty) of prediction about non-social cue $\widehat{\sigma}_{1\ card}$ and D) social cue $\widehat{\sigma}_{1\ gaze.}$ E) The red trajectory shows the posterior expectation of the blue card to be correct. The true trial outcomes with respect to the blue card (blue correct=1; green correct=0) are shown in dark red dots and the responses with respect to the card (blue card=1; green card=0) shown in light red dots. F) The blue trajectory shows the posterior expectation of the social advice to be correct. The true trial outcomes with respect to the gaze (correct=1; incorrect=0) are shown in dark blue dots and the responses with respect to the gaze (follow=1; not follow=0) are shown in light blue dots. Green dots marked missed trials.

2.8 Eyetracking Data Acquisition and Analysis

Eye movement data was acquired employing an infrared pupil-corneal reflection-based eye-tracking system (Eyelink 1000 Plus, SR Research, Osgoode, ON, Canada), which was connected to an MR compatible fibre-optic camera head. The camera head consisted of a 75 mm lens and an MR- compatible LED Illuminator. A first-surface reflecting mirror was attached to the scanner head coil to reflect participants' eye movements. The distance between mirror and eye-tracker was 125 cm and the distance between eyes and monitor was 240 cm. We used a nine-point calibration to map the gaze position onto screen coordinates and we acquired data using a sampling rate of 2000 Hz. Preprocessing of eye tracking data was performed using Matlab (Version 2017a; www.mathworks.com). We segmented fixations during the choice phase starting from the point of the advice until the response of the participant. We also calculated mean fixation points during the inter-trial interval (ITI). Due to the long operating distance between eyes and monitor in the scanner, we observed a shift in fixation data, which was different for all participants. We calculated a shift distance in the x and y coordinates for each participant by subtracting the mean measured fixation points during the ITI's from the coordinates of the fixation cross that was presented during the ITI. This shift value for both coordinates was then applied to the segmented fixation points of the decision and outcome phase.

In order to investigate the relationship between ζ and the gaze data further, we used a general linear model approach similar to the one employed in the fMRI analyses: We created participant-specific fixation heatmaps for each trial (768x1024x120) for the choice phase as well as for the outcome phase. When generating the heatmaps, we smoothed the fixation maps using a Gaussian kernel with mu of fixation's Cartesian coordinate and SD of 1° corresponding to a full-width-at-half-maximum of approximately 2.35° (Lahnakoski et al., 2014). We further defined pixel-wise GLMs to analyse those regions of the screen where the number of fixations correlate with the social weighting factor ζ.

Furthermore, in order to incorporate fixation data into our GLM model, we calculated the proportion of face fixations during the decision phase. For this, we counted fixation points falling onto the region of the screen where the face was presented and fixation points falling on all remaining parts of the screen. We then divided the number of fixations points from the rest of the screen by the number of fixation points falling on the face. In all eye-tracking analyses, 6 participants had to be discarded from further analysis due to blurred corneal reflection signals.

3. Results

3.1 Bayesian Model Comparison & Selection

Random effects BMS revealed a clear superiority for the three-level HGF in combination with a response model in which decision noise is a combination of the log-volatility for both gaze and card combined with participant-specific log-volatility for card $\hat{\mu}_{3,card}$ and participant-specific decision noise β (XP= 0.9370; PXP= 0.627; EXP_P = 0.464; Table 1). Therefore, we used this model for all subsequent analyses. Mean parameter estimates can be seen in the appendix (Tab A. 3).

3.2 Simulations

While keeping the perceptual model parameters fixed at the prior values, we simulated inferred choice probabilities (in gaze space (Equation 9) and in card space) of agents with variable ζ values to investigate how this parameter will affect choice probabilities with regard to the social information (Figure 3A) and the non-social information (Figure 3B) respectively. The simulations show that ζ represents a relative sensitivity parameter for the social input over the non-social input such that high ζ values mean that the integrated belief is characterized by an increased sensitivity of the social information (gaze correct vs. gaze wrong) and at the same time a decreased sensitivity, i.e. increased stochasticity, with regard to the non-social information (blue card vs. green card correct).

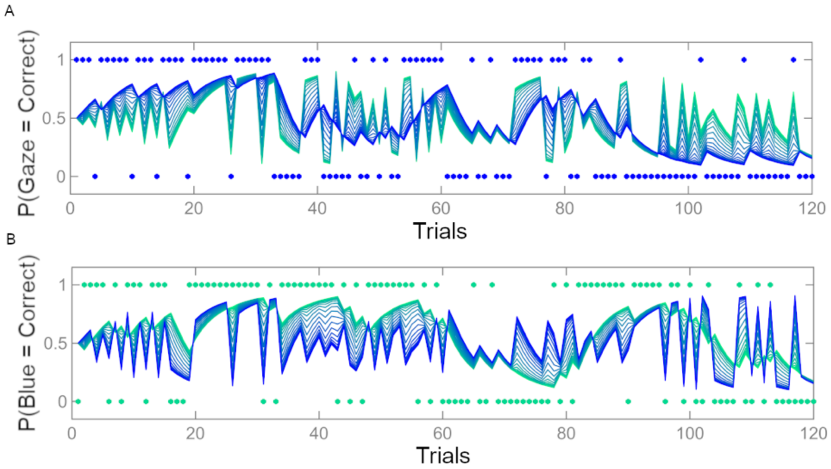


Figure 3: Simulation for an agent with same perceptual parameters but different social cue weighting. The plot shows A) the different probability trajectories for taking the advice (p (y=1 | b) with varying ζ values (highest values (log(5) coded in blue, lowest values (log(-5) coded in light colours) and B) different probability trajectories for taking the blue card (p(y=1 | b) with varying ζ values (highest values (log(5) coded in blue, lowest values (log(-5) coded in light colours). The actual input of the gaze (1= correct; 0=incorrect) is shown in blue in A) and the input of the card on a given trial (1= blue; 0=green) is shown in green in B).

3.3 Behavioural statistics: Advice-taking & fixation behaviour

We found that the social weighting factor ζ was significantly correlated with subjective reports of having used the gaze during the task ($r_S(48) = 0.453$, $p = 0.001$) and the subjective report of finding the gaze helpful ($r_S(48) = 0.292$, $p = 0.044$) (Figure 4A,B). The social weighting factor ζ was positively correlated with the proportion of trials in which the gaze was followed (r_S (48) = 0.487, $p < 0.001$) (Figure 4C). The same was the case for the subjective report of using the gaze (r_S (48) = 0.449, $p = 0.001$). Furthermore, when looking at advice-taking behaviour, the repeated measures ANOVA revealed a main effect of task phase ($F(2,96) = 57.050$, $p < 0.001$, η^2=0.543) showing that participants' advice-taking behaviour varied with the probability by which the gaze was giving a helpful advice. Post-hoc t-tests showed that participants followed the advice significantly more often in the high-accuracy phase (80%) compared to the volatile phase ($t(50)$= 7.357, $p < 0.001$, d =1.04). During the low-accuracy phase (20%) participants chose the advice significantly less compared to the volatile ($t(50)$= 2.911, $p = 0.016$, $d = 0.41$) and compared to the high accuracy phase ($t(50)$= 8.340, $p < 0.001$, $d = 1.179$).

There was a main effect of the covariate ζ ($F(1,48) = 17.54$, $p < 0.001$, η^2=0.268) and a significant interaction between the covariate ζ and the magnitude of the effect of task phase on behaviour ($F(2,96) = 6.832$, $p = 0.002$, η^2=0.125) indicating that participants with a higher ζ are more sensitive to the social cue probability. Furthermore, the GLM analysis of the fixation data revealed that fixation points falling on the face area of the social stimulus ($p <$ 0.001, uncorrected) during choice phase were significantly correlated with ζ (Figure 4D).

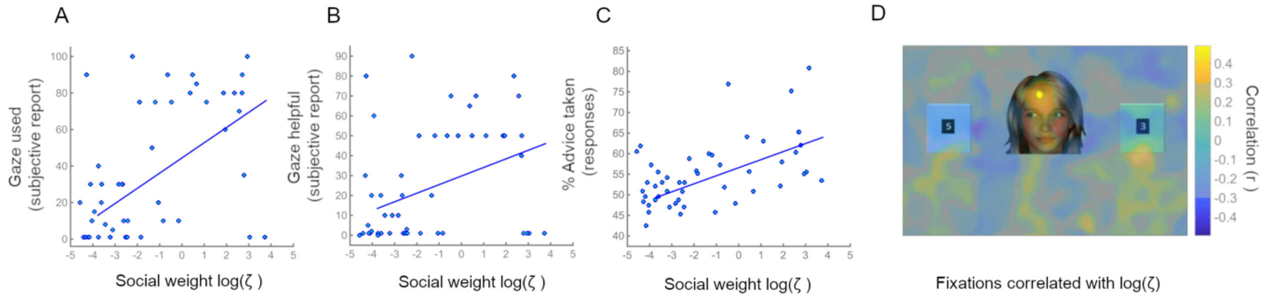


Figure 4 A) Association between estimated values of computational parameter ζ and subjective reports of having used the gaze and B) finding it helpful during decision-making. C) Association between ζ and % of trials where gaze was followed. D) Mean proportion of fixations on face area during all trials and ζ; Pixel-wise analysis of

smoothed fixation data revealed that ζ is correlated with the time people spend looking at the face ($p<0.001$, uncorrected) during the choice phase of the trials.

3.4 fMRI results

Social and non-social prediction and precision during decision-making

During the choice phase of the task, the subjective predicted advice accuracy $\hat{\mu}_{1,gaze}^{(t)}$ correlated with activity in the right and left inferior temporal gyri, left and right inferior parietal lobule, left and right precentral gyri, right postcentral gyrus, left and right superior frontal gyrus, left and right fusiform gyri, and the right putamen, superior orbital gyrus and pallidum (Figure 5 and Table 2). Self-reports of having used the gaze during decision-making were associated with higher activity related to $\hat{\mu}_{1,gaze}^{(t)}$ in the right rectal gyrus, right and left putamen and insula (Figure 6 and Table 3) across participants. Differences in activation strength as a function of ζ were associated with activity in the right inferior occipital gyrus (Table A.4). Significant clusters were neither found for the correlation with $1-\hat{\mu}_{1,gaze}^{(t)}$ (the subjective predicted probability of a misleading gaze) nor for the variance of the prediction $\hat{\sigma}_{1,gaze}^{(t)}$ and $1-\hat{\sigma}_{1,gaze}^{(t)}$. Results for $\hat{\mu}_{1,gaze}^{(t)}$ when estimated together with $\hat{\sigma}_{1,gaze}^{(t)}$ and $\hat{\sigma}_{1,card}^{(t)}$ in one GLM can be seen in Table A.5 & A.6.

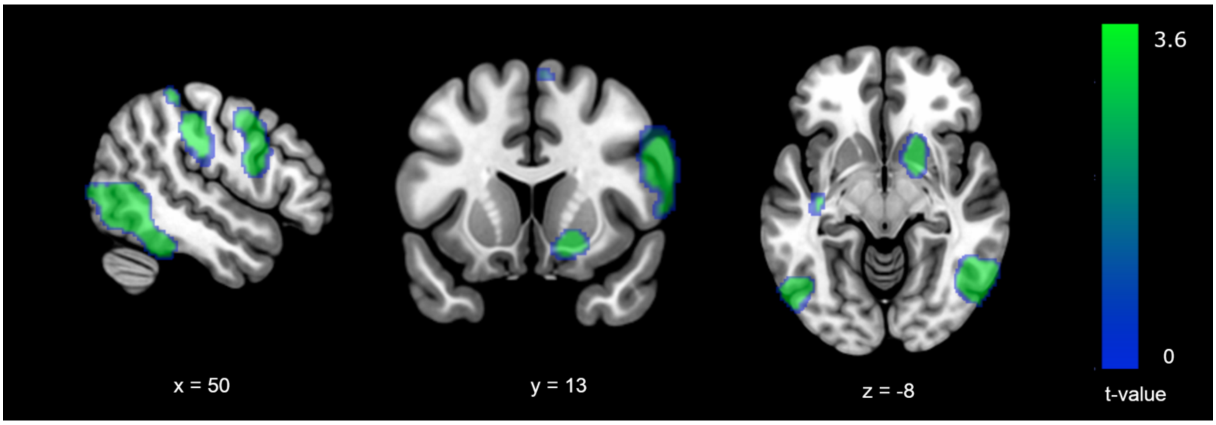


Fig. 5 fMRI results for predicted accuracy of advice ($\hat{\mu}_{1,gaze}^{(t)}$) during the choice phase of the task. Cluster-forming threshold: $p<0.001$, cluster-level threshold $p< 0.05$, FWE corrected. [x y z] coordinates refer to the MNI coordinates of the respective slices. See Table 2 for further information on cluster extents and peak voxel coordinates.

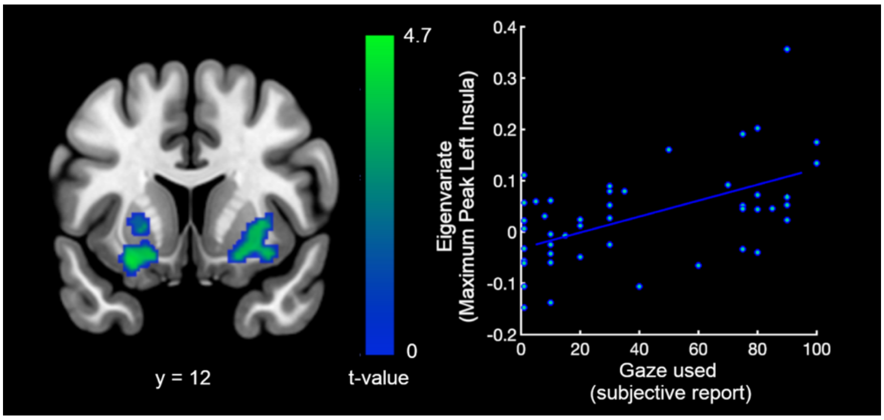


Fig. 6 The contrast in the left shows brain areas showing differential responses to $\hat{\mu}_{1,gaze}^{(t)}$as a function of the subjective report of having used the gaze during decision-making. The right plot depicts the correlation between the subjective report and the highest peak in the insula. Cluster-forming threshold: p<0.001, cluster-level threshold p< 0.05, FWE corrected. The Y coordinate refer to the MNI coordinate of the respective slice. See Table 3 for further information on cluster extents and peak voxel coordinates.

In the choice phase of the task, the negative contrast on the variance of the belief about the winning card colour (1-$\widehat{\sigma}_{1,card}^{(t)}$) correlated with the right insula and right rolandic operculum (Fig. 7 and Table 4). Neither the $-\zeta$ (computational non-social weight) nor $-$ *Question3* (subjective non-social weight), were correlated with brain activity related to 1-$\widehat{\sigma}_{1,card}^{(t)}$. No significant clusters were found for the positive contrast ($\widehat{\sigma}_{1,card}^{(t)}$).

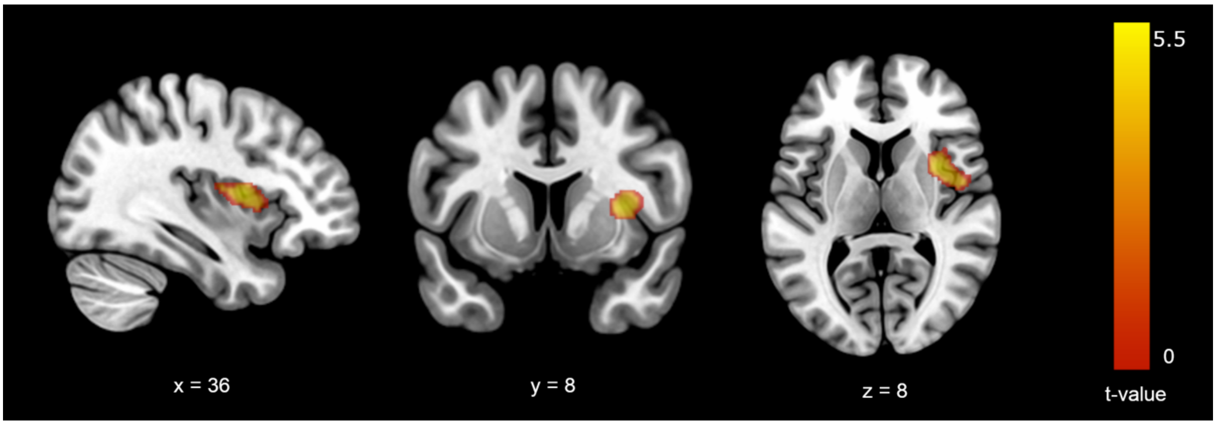


Fig. 7. Significant clusters for the negative contrast of the variance of the prediction of the winning card colour ($\widehat{\sigma}_{1,card}^{(t)}$) during the choice phase of the task. Cluster-forming threshold: p< 0.001 uncorrected, cluster-level threshold p < 0.05, FWE corrected. [x y z] coordinates refer to the MNI coordinates of the respective slices. See Table 4 for further information on cluster extents and peak voxel coordinates.

Social and non-social prediction error during outcome

For negative social prediction errors ($\delta_{1gaze}^{(t)} < 0$, i.e., gaze misleading) during wrong choice outcomes, we observed significant activations in the right inferior frontal gyrus, right insula, rolandic operculum and left posterior medial frontal gyrus (Fig 8 and Table 5). No significant activations were found for negative social prediction errors during correct choice outcomes or when evaluating both correct and wrong choices. The analyses looking at the positive prediction error signals ($\delta_{1gaze}^{(t)} > 0$, i.e., gaze helpful) revealed significant activations in the right lingual gyrus and middle occipital gyrus, but only in correct choice outcomes (Table 6).

When we directly compared negative prediction errors during incorrect outcomes against negative prediction errors during correct outcomes we found the same activation in the right insula, rolandic operculum and left posterior medial frontal gyrus gyrus as when evaluating negative social prediction errors against baseline during incorrect outcomes. When we directly compared positive prediction error signals during correct outcomes against positive prediction errors during incorrect outcomes, we also found the same activation in the right lingual gyrus and middle occipital gyrus as when evaluating positive social prediction errors against baseline during correct outcomes.

Comparing negative with positive social prediction errors, we found the same activation in the right inferior frontal gyrus, right insula, rolandic operculum and left posterior medial frontal gyrus but only when evaluating incorrect outcomes. The activation was found in the same regions as when evaluating negative social prediction errors against baseline during wrong outcomes.

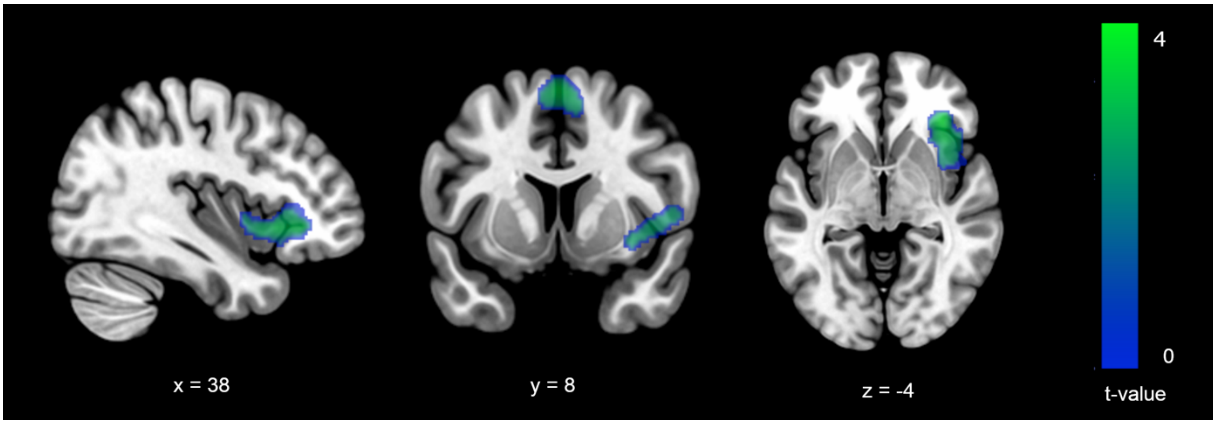


Fig. 8 Neural correlates of ($\delta_{1gaze}^{(t)}$) during wrong choice outcomes. The negative contrast on the parametric modulator of the outcome phase reflects BOLD activity in regions correlated with negative social prediction errors. Cluster-forming threshold: p<0.001, cluster-level threshold p< 0.05, FWE corrected. [x y z] coordinates refer to the MNI coordinates of the respective slices. See Table 5 for further information on cluster extents and peak voxel coordinates.

Next, we looked at the absolute prediction error of the advice ($\delta_{1card}^{(t)}$), signalling the surprise about the cue colour. When the social cue was correct, we found significant bilateral activations in the posterior medial frontal gyri, anterior and middle cingulate cortex and insula (Fig 9, Table 7). When looking at the modulation of $\delta_{1card}^{(t)}$ during all outcomes irrespective of advice accuracy, only the cluster in the posterior-medial, superior frontal gyrus and middle cingulate cortex and the cluster in the left insula was significant (Table 7). When the social cue was incorrect, no significant clusters were found for $\delta_{1card}^{(t)}$. The results for the negative contrast on $\delta_{1card}^{(t)}$ looking at activity correlated with a decrease in surprise about the winning card colour can be seen in Table A. 7.

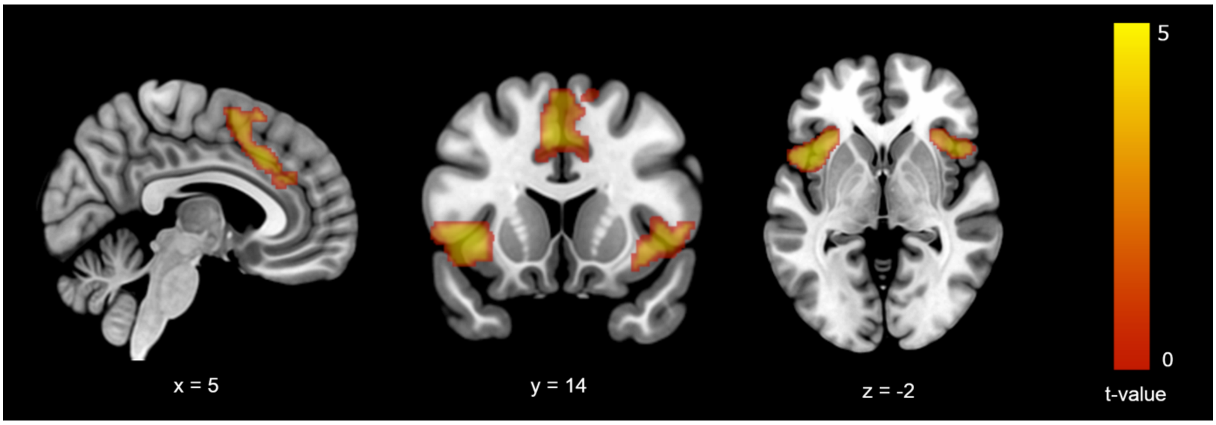


Fig. 9 Neural correlates of absolute learning signal ($\delta_{1card}^{(t)}$). The positive contrast on the parametric modulator reflects BOLD activity in regions correlated with amount of surprise about the accuracy of the card colour when social cue was correct. Cluster-forming threshold: p<0.001, cluster-level threshold p< 0.05, FWE corrected. [x y z] coordinates refer to the MNI coordinates of the respective slices. See Table 7 for further information on cluster extents and peak voxel coordinates.

4. Discussion

In this study, we used model-based fMRI to uncover the neural mechanisms of inference on social and non-social cues during a probabilistic learning task using a three-level hierarchical Bayesian model describing parallel learning. Furthermore, we assessed individual differences in the relative weight granted to social over non-social information during the task and demonstrated that the estimated values of the corresponding parameter accord with model-agnostic equivalents such as subjective reports and eye gaze behaviour during the task. In addition, we showed that the weight on social information during decision-making correlates with individual differences in brain activation during decision-making, in particular in the putamen and insula.

4.1 Social and non-social prediction error activations

Negative social prediction errors ($\delta_{1gaze}^{(t)} < 0$, i.e., gaze misleading) during wrong choices recruited the right anterior insula, as well as the right inferior frontal gyrus and the left posterior-medial frontal gyrus. For correct choices these activations were absent, suggesting that the deception of the social cue was not relevant when participants succeeded in selecting the winning card on a given trial.

Activation in the anterior insula in response to negative social prediction errors is in line with insula activity in response to misleading advice in a previous study of explicit mentalizing (Diaconescu et al., 2017), as well as unreciprocated cooperation in the trust game (King-Casas et al., 2008; Rilling et al., 2008), social exclusion (Eisenberger et al., 2003) and to (negative) surprise about the expected offer of a confederate in a fairness game (Xiang, Lohrenz, & Montague, 2013). These findings support the notion that the anterior insula plays an important role in tracking risk in uncertain environments (Bossaerts, 2010; d'Acremont, Lu, Li, Van der Linden, & Bechara, 2009). In particular, the right anterior insula has been found to be involved in the integration of (arousing) interoceptive states into decision-making, potentially by signalling aversive events that are to be avoided in the future (Rilling, King-Casas, & Sanfey, 2008). In our study, participants did not know if and to what extent the social cue will provide them helpful or misleading advice. The activity in the insula and inferior frontal gyrus to negative social prediction errors (i.e. misleading advice) was only observed in trials in which participants did not receive the reward. In other words, the insula/inferior frontal gyrus activation signalled occasions where the participant should not have followed the gaze.

We also found significant correlations with non-social prediction errors $\delta_{1card}^{(t)}$ in the left and right insula, a pattern resembling prediction error activation in a sensory learning paradigm (Iglesias et al., 2013), which underlines the insula's role in error monitoring irrespective of the domain of learning (Diaconescu et al., 2017).

For positive social prediction errors (gaze more helpful than predicted) during correct outcomes, we found activity in the right occipital and lingual gyrus but not in reward-associated areas as reported by others (Biele et al., 2011; Delgado et al., 2005; Fareri, Chang, & Delgado, 2012; Fouragnan et al., 2013). This may reflect the directing of visual attention towards relevant, in our case, social stimuli. Indeed, reward learning signals were previously also found in the occipital cortex by Payzan-LeNestour, Dunne, Bossaerts, & O'Doherty (2013).

In the present study, social prediction errors did not significantly activate brain regions that have been associated with mentalization, such as the TPJ or the dmPFC (Behrens et al., 2008;

Diaconescu et al., 2017; Koster-Hale et al. 2017) or that have been associated with observational learning such as the ACCg (Apps et al., 2015, 2016; Lockwood et al., 2015). A crucial difference between the present and other social learning studies is that our study did not involve instruction with respect to an opponent or confederate. Instead, we merely presented the computer-generated face because we wanted to investigate the spontaneous integration of social information into decision making. Indeed, a subgroup of our participants claimed not to have used the social information during the task. Possibly, these participants concentrated more on the non-social feedback to predict the outcome, relying less on social feedback to adapt their behaviour, thus reducing statistical power to detect effects of social inference in the group analysis.

4.2 Social and non-social prediction and precision

We found that the belief about the social cue, i.e. the inferred probability of the gaze to give a correct advice ($\hat{\mu}_{1,gaze}^{(t)}$), was associated with activity in the inferior temporal gyri, inferior and superior parietal lobule as well as parts of the striatum including the right putamen and pallidum. The striatum's involvement in tracking the belief about the accuracy of social advice during choice accords with earlier findings regarding the role of this region in encoding the value of social interaction partners (Báez-Mendoza & Schultz, 2013; Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008; Delgado et al., 2005; King-Casas et al., 2005; Rilling et al., 2008) and of the non-social aspects of a learning environment (cf. O'Doherty, 2004).

The present results suggest that the magnitude of BOLD activity related to advice accuracy in the putamen and anterior insula may be modulated as a function of individual differences in employing the social cue during decision-making. Specifically, the recruitment of the putamen and insula was more pronounced for participants that integrated the social cue into their decision-making, as indicated by subjective reports. Activity changes in the insula that correlate with advice accuracy during choice are in line with a previous finding of insula activity correlating with the predicted value of the action of another person (Apps et al., 2015).

Our finding that putamen and insula activities were correlated with increased weighting of social information needs to be seen in light of a limitation of the current study: we did not have a non-social control condition, for instance in form of an arrow pointing to one of the cards. Therefore, we cannot fully determine whether individual differences in social cue weighting associated with insula and putamen activity can be attributed to purely social or more general learning processes. In fact, co-activation of putamen and insula has previously been found in non-social cueing tasks (Hopfinger, Buonocore, & Mangun, 2000). Remarkably however, these regions show significantly stronger activations for directional gaze cues compared to arrows in a spatial cueing task in healthy participants (Greene et al., 2011).

These findings raise the potential of our method for studying aberrant social inference in psychiatric disorders (Diaconescu, Hauke, & Borgwardt, 2019; Frith, 2004), which is often associated with deficits in automatic but not explicit integration of social cues (Callenmark, Kjellin, Ronnqvist, & Bolte, 2014; Senju, Southgate, White, & Frith, 2009). Specifically, patients with schizophrenia have a tendency to over-attribute the meaning and salience of social signals (Diaconescu, Hauke, & Borgwardt, 2019; Frith, 2004). It would be interesting to investigate whether this would be reflected in processing abnormalities in the insula and putamen.

Interestingly, while we found significant activations in the right insula correlating negatively with uncertainty about the winning card colour, we did not find differential activity in the insula as a function of non-social cue weighting ($-\zeta$). While we did not find significant activations with regard to uncertainty about the social cue, we found that fixation frequency on the face during choice, which may in itself reflect the degree of decision uncertainty (Brunyé & Gardony, 2017), was correlated with activations in the superior temporal gyrus (at a less conservative statistical threshold). This is in line with this region's role in mentalization and suggests that these processes are triggered in the absence of explicit instructions to mentalize.

5. Conclusions

The present study used model-based fMRI to demonstrate commonalities and differences in the neural mechanisms of social and non-social cue integration during learning and decision-making. While activations related to the non-social cue were associated with activity change in the middle and anterior cingulate and insula, negative social prediction errors additionally extended into the inferior frontal gyrus. During decision-making, tracking the uncertainty of the non-social cue was associated with activity change in the insula, while tracking the probabilistic accuracy of the social cue showed activity in the inferior temporal gyrus, putamen and pallidum, regions known for their relevance in reward-based processing. The putamen and the insula showed activity as a function of individual differences in weighting the social cue during decision-making. Our findings demonstrate the usefulness of model-based fMRI for the study of the spontaneous use of social cues in learning and decision-making, and they provide evidence for the involvement of specific components of the basal ganglia in these processes.

Acknowledgements

We would like to thank Ines Eidner for helping with the data collection and Alana Darcher for her help in pre-processing the eye tracking data.

Funding

This work was supported by a grant to Leonhard Schilbach for an independent Max Planck Research Group. AOD acknowledges the support by the Swiss National Foundation Ambizione PZ00P3_167952.

References

Apps, M. A. J., Lesage, E., & Ramnani, N. (2015). Vicarious reinforcement learning signáis when instructing others. *Journal of Neuroscience*, *35*(7), 2904–2913. https://doi.org/10.1523/JNEUROSCI.3669-14.2015

Apps, M. A. J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. *Neuron*, *90*(4), 692–707. https://doi.org/10.1016/j.neuron.2016.04.018

Báez-Mendoza, R., & Schultz, W. (2013). The role of the striatum in social behavior. *Frontiers in Neuroscience*, *7*(7 DEC), 1–14. https://doi.org/10.3389/fnins.2013.00233

Bagby, R. M., Taylor, G. J., & Parker, J. D. A. (1994). The twenty-item Toronto Alexithymia scale-II. Convergent, discriminant, and concurrent validity. *Journal of Psychosomatic Research*, *38*(1), 33–40. https://doi.org/10.1016/0022-3999(94)90006-X

Baron-cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). *The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism*. *42*(2), 241–251. https://doi.org/10.1017/S0021963001006643

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/10.1023/A:1005653411471

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin Shapes the Neural Circuitry of Trust and Trust Adaptation in Humans. *Neuron*, *58*(4), 639–650. https://doi.org/10.1016/j.neuron.2008.04.009

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. https://doi.org/10.1038/nature07538

Biele, G., Rieskamp, J., Krugel, L. K., & Heekeren, H. R. (2011). The Neural basis of following advice. *PLoS Biology*, *9*(6). https://doi.org/10.1371/journal.pbio.1001089

Bossaerts, P. (2010). Risk and risk prediction error signals in anterior insula. *Brain Structure & Function*, *214*(5–6), 645–653. https://doi.org/10.1007/s00429-010-0253-1

Brunyé, T. T., & Gardony, A. L. (2017). Eye tracking measures of uncertainty during perceptual decision making. *International Journal of Psychophysiology*, *120*(April), 60–68. https://doi.org/10.1016/j.ijpsycho.2017.07.008

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14431–14436. https://doi.org/10.1073/pnas.1003111107

Callenmark, B., Kjellin, L., Ronnqvist, L., & Bolte, S. (2014). Explicit versus implicit social cognition testing in autism spectrum disorder. *Autism*, *18*(6), 684–693. https://doi.org/10.1177/1362361313492393

d'Acremont, M., Lu, Z. L., Li, X., Van der Linden, M., & Bechara, A. (2009). Neural correlates of risk prediction error during reinforcement learning in humans. *NeuroImage*, *47*(4), 1929–1939. https://doi.org/10.1016/j.neuroimage.2009.04.096

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the Observer (I): Meta-Bayesian Models of Learning and Decision-Making. *PLoS ONE*, *5*(12), e15554. https://doi.org/10.1371/journal.pone.0015554

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453. https://doi.org/10.3758/CABN.8.4.429

DeBerker, A. O. De, Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*, *7*, 1–11. https://doi.org/10.1038/ncomms10996

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611–1618. https://doi.org/10.1038/nn1575

Diaconescu, A., Hauke, D. J., & Borgwardt, S. (2019). Models of persecutory delusions: a mechanistic insight into the early stages of psychosis. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-019-0427-z

Diaconescu, A., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, *12*(4), 618–634. https://doi.org/10.1093/scan/nsw171

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, (November 2016), nsw171. https://doi.org/10.1093/scan/nsw171

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*(5643), 290–292.

https://doi.org/10.1126/science.1089134

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015). Computational Substrates of Social Value in Interpersonal Collaboration. *Journal of Neuroscience*, *35*(21), 8170–8180. https://doi.org/10.1523/jneurosci.4775-14.2015

Fareri, Dominic S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience*, *6*(OCT), 1–17. https://doi.org/10.3389/fnins.2012.00148

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational Priors Magnify Striatal Responses to Violations of Trust. *Journal of Neuroscience*, *33*(8), 3602–3611. https://doi.org/10.1523/jneurosci.3086-12.2013

Frith, C. (2004). Schizophrenia and theory of mind. *Psychological Medicine*, *34*, 385–389. https://doi.org/10.1017/S0033291703001236

Gooding, D. C., & Pflum, M. J. (2014). The assessment of interpersonal pleasure: Introduction of the Anticipatory and Consummatory Interpersonal Pleasure Scale (ACIPS) and preliminary findings. *Psychiatry Research*, *215*(1), 237–243. https://doi.org/10.1016/j.psychres.2013.10.012

Greene, D. J., Colich, N., Iacoboni, M., Zaidel, E., Bookheimer, S. Y., & Dapretto, M. (2011). Atypical neural networks for social orienting in autism spectrum disorders. *NeuroImage*, *56*(1), 354–362. https://doi.org/10.1016/j.neuroimage.2011.02.031

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233–1235. https://doi.org/10.1038/nn.4080

Hopfinger, J. B., Buonocore, M. H., & Mangun, G. R. (2000). The neural mechanisms of top-down attentional control. *Nature Neuroscience*, *3*(3), 284–291. https://doi.org/10.1038/72999

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., denOuden, H. E. M., & Stephan, K. E. (2013). Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron*, *80*(2), 519–530. https://doi.org/10.1016/j.neuron.2013.09.009

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., Ouden, H. E. M. Den, & Stephan, K. E. (2013). *Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning*. 519–530.

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *Npj Science of Learning*, *2*(1), 1–9. https://doi.org/10.1038/s41539-

017-0009-2

Kasper, L., Bollmann, S., Diaconescu, A. O., Hutton, C., Heinzle, J., Iglesias, S., … Stephan, K. E. (2017). The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. *Journal of Neuroscience Methods*, *276*, 56–72. https://doi.org/10.1016/j.jneumeth.2016.10.019

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, R. (2005). Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science*, *308*(5718), 78–83. https://doi.org/10.1126/science.1108062

King-Casas, Brooks, Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science (New York, N.Y.)*, *321*(5890), 806–810. https://doi.org/10.1126/science.1156902

Kühner, C., Bürger, C., Keller, F., & Hautzinger, M. (2006). Reliabilität und Validität des revidierten Beck-Depressionsinventars (BDI-II)Reliability and validity of the Revised Beck Depression Inventory (BDI-II). *Der Nervenarzt*, *78*(6), 651–656. https://doi.org/10.1007/s00115-006-2098-7

Lahnakoski, J. M., Glerean, E., Jääskeläinen, I. P., Hyönä, J., Hari, R., Sams, M., & Nummenmaa, L. (2014). Synchronous brain activity across individuals underlies shared psychological perspectives. *NeuroImage*, *100*, 316–324. https://doi.org/10.1016/j.neuroimage.2014.06.022

Linden, M., Lischka, A.-M., Popien, C., & Golombek, J. (2007). Der multidimensionale Sozialkontakt Kreis (MuSK) – ein Interviewverfahren zur Erfassung des sozialen Netzes in der klinischen Praxis. *Zeitschrift Für Medizinische Psychologie*, *16*(3), 135–143. Retrieved from http://iospress.metapress.com/content/00NQP0M6X7261627

Lockwood, P. L., Apps, M. A. J., Roiser, J. P., & Viding, E. (2015). Encoding of vicarious reward prediction in anterior cingulate cortex and relationship with trait empathy. *Journal of Neuroscience*, *35*(40), 13720–13727. https://doi.org/10.1523/JNEUROSCI.1703-15.2015

Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences*, *113*(35), 9763–9768. https://doi.org/10.1073/pnas.1603198113

Lockwood, P. L., & Klein-Flüggea, M. (2019). *Computational modelling of social cognition and behaviour – a reinforcement learning primer*. 1–28.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., &

Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*(November), 825. https://doi.org/10.3389/fnhum.2014.00825

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*(May), 1–20. https://doi.org/10.3389/fnhum.2011.00039

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology*, *14*(6), 769–776. https://doi.org/10.1016/j.conb.2004.10.016

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making. *Annual Review of Psychology*, *68*(1), 73–100. https://doi.org/10.1146/annurev-psych-010416-044216

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2013). The Neural Representation of Unexpected Uncertainty during Value-Based Decision Making. *Neuron*, *79*(1), 191–201. https://doi.org/10.1016/j.neuron.2013.04.037

Rescorla, R. A., & Wagner, A. R. (1972). *A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement*. 1–18. https://doi.org/10.1101/gr.110528.110

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). NeuroImage Bayesian model selection for group studies — Revisited. *NeuroImage*, *84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Rilling, J. K., King-Casas, B., & Sanfey, A. G. (2008a). The neurobiology of social decision-making. *Current Opinion in Neurobiology*, *18*(2), 159–165. https://doi.org/10.1016/J.CONB.2008.06.003

Rilling, J. K., King-Casas, B., & Sanfey, A. G. (2008b). The neurobiology of social decision-making. *Current Opinion in Neurobiology*, *18*(2), 159–165. https://doi.org/10.1016/j.conb.2008.06.003

Ruff, C. C., & Fehr, E. (2014). *The neurobiology of rewards and values in social decision making*. (July). https://doi.org/10.1038/nrn3776

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Senju, A., Southgate, V., White, S., & Frith, U. (2009). *Mindblind Eyes : An Absence of Asperger Syndrome*. *219*(August), 883–885. https://doi.org/10.1126/science.1176170

Sevgi, M., Diaconescu, A. O., Henco, L., Tittgemeyer, M., & Schilbach, L. (2020). Social Bayes: Using Bayesian Modeling to Study Autistic Trait–Related Differences in Social Cognition. *Biological Psychiatry*, *87*(2), 185–193. https://doi.org/10.1016/j.biopsych.2019.09.032

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Sutton, R. S. (1992). Gain Adaptation Beats Least Squares? *Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems*, 161–166. Retrieved from papers://d471b97a-e92c-44c2-8562-4efc271c8c1b/Paper/p596

Wittmann, M. K., Lockwood, P. L., & Rushworth, M. F. S. (2018). Neural Mechanisms of Social Cognition in Primates. *Annual Review of Neuroscience*, *41*(1), 99–118. https://doi.org/10.1146/annurev-neuro-080317-061450

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, *33*(3), 1099–1108. https://doi.org/10.1523/JNEUROSCI.1642-12.2013

No part of the study procedures or analyses were pre-registered prior to the research being conducted.

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

The conditions of our ethical approval do not permit public archiving or peer-to-peer sharing of individual raw data. The data supporting the conclusions of this article are therefore not available to any individual outside the author team under any circumstances.

All digital materials associated with this experiment, presentation code, and analysis scripts will be made available.

Tables

Table 1. Bayesian Model Selection results. Posterior model probabilities (EXP_R) and Protected Exceedance Probabilities (PXP).

	Model 1	Model 2	Model 3	Model 4	Model 4	Model 6
EXP_R	0.464	0.098	0.077	0.031	0.289	0.042
PXP	0.627	0.067	0.067	0.067	0.105	0.067
XP	0.937	0	0	0	0.063	0

Table 2. fMRI results for predicted accuracy of advice ($\hat{\mu}^{(t)}_{1.gaze}$) during the choice phase.

		Cluster		MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
R Inferior Temporal Gyrus	0	1749	7.9	52	-60	-6
R Fusiform Gyrus			4.14	40	-72	-18
R Middle Occipital Gyrus			4.01	50	-82	4
L SupraMarginal Gyrus	0	1057	6.25	-58	-24	36
L Inferior Parietal Lobule			4.4	-54	-36	50
R Precentral Gyrus	0	4401	6.15	58	10	30
L Precentral Gyrus			5.32	-34	-10	58
R Superior Frontal Gyrus			5.14	26	-6	68
L Posterior-Medial Frontal			4.8	-8	-4	68
L Superior Frontal Gyrus			4.75	-22	-8	72
L Inferior Parietal Lobule			4.74	-34	-42	50
R Superior Frontal Gyrus			4.73	20	4	72
R Postcentral Gyrus	0	1424	5.91	54	-22	34
R SupraMarginal Gyrus			5.62	62	-16	28
R Inferior Parietal Lobule			4.03	44	-34	48
L Inferior Temporal Gyrus	0.001	524	5.77	-50	-68	-8
L Middle Temporal Gyrus			3.29	-56	-58	2
White Matter	0.004	421	5.04	16	6	-12
R Putamen			4.32	18	14	-10
R Pallidum			4.08	22	2	0
R Superior Orbital Gyrus			3.95	18	22	-18
L Inferior Temporal Gyrus	0.028	274	4.39	-40	-28	-26
L Fusiform Gyrus			4.23	-38	-32	-28
L Inferior Temporal Gyrus			4.21	-44	-24	-20
L Cerebelum (VI)			3.93	-32	-40	-28

Table 3. Neural correlates of differential responses to $\hat{\mu}^{(t)}_{1.gaze}$ as a function of the subjective report of having used the gaze during decision making.

		Cluster		MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
L Insula Lobe	281	281	4.78	-26	12	-16
L Putamen			4.04	-22	16	0
R Rectal Gyrus	234	234	4.65	20	18	-12

R Putamen			3.83	30	10	0
R Insula Lobe			3.36	36	6	12

Table 4. fMRI results for the negative contrast on the predicted variance of the winning card colour $\hat{\sigma}_{1.card}^{(t)}$ during the choice phase.

		Cluster		MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
R Insula Lobe	0.003	381	5.51	36	6	10
R Rolandic Operculum			4.68	46	-2	14

Table 5. fMRI results for negative social prediction error ($\delta_{1gaze}^{(t)} < 0.$ i.e. gaze misleading) during wrong choices.

		Cluster		MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
R Inferior Frontal Gyrus (p. Orbitalis)	0	769	5.9	36	32	-4
R Insula Lobe			4.95	36	22	-4
R Inferior Frontal Gyrus (p. Triangularis)			4.82	50	28	2
R Rolandic Operculum			3.84	52	8	4
L Posterior-Medial Frontal	0.009	312	4.87	0	4	64

Table 6. fMRI results for positive social prediction error ($\delta_{1gaze}^{(t)} > 0.$ i.e. gaze helpful) during correct choices.

		Cluster		MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
R Lingual Gyrus		499	4.59	14	-98	-8
R Middle Occipital Gyrus			3.66	36	-96	0

Table 7. fMRI results for $|\delta_{1card}^{(t)}|$ during outcome phases where advice was correct.

		Cluster		MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
Advice correct						
L Insula Lobe	0	568	5.54	-42	14	-2
R Middle Cingulate Cortex	0	1020	5.39	8	20	38
L Posterior-Medial Frontal			5.24	-4	12	48
L Middle Cingulate Cortex			5.1	-2	20	38
R Posterior-Medial Frontal			4.1	8	8	54
R Anterior Cingulate Cortex			3.78	6	30	26
L Anterior Cingulate Cortex			3.55	2	38	26
R Inferior Frontal Gyrus (p. Orbitalis)	0.002	422	5.32	34	24	-8
R Insula Lobe			5.16	42	18	-4
All outcomes						

L Insula Lobe	0.032	225	5.11	-44	12	-4
L Posterior-Medial Frontal	0.001	492	4.81	-4	12	48
R Superior Frontal Gyrus			4.08	14	4	74
R Middle Cingulate Cortex			3.79	8	20	36

APPENDIX for “Bayesian modelling captures inter-individual differences in social belief computations in the putamen and insula”

1. Methods

Table A. 1 Descriptive statistics of post-experimental questions all participants. After the experiment subjects were asked to rate the difficulty of the task from 0 - easy to 100 - very difficult (Q1). how helpful they found the gaze during decision making from 0 – not helpful at all to 100 – very helpful (Q2) and how much they took the gaze into account from 0 – not at all to 100 – very much (Q3).

	Q1	Q2	Q3
Mean	38.19	26.48	38.10
Std. Deviation	28.17	28.29	35.27
Minimum	1	0	1
Maximum	100	90	100
Missing values	3	2	2

Tab. A. 2. Details on prior configurations for all perceptual models (HGF, ST-K1 & RW) and response model.

	Level 1		Level 2		Level 3	
HGF	Prior Mean	Prior Variance	Prior Mean	Prior Variance	Prior Mean	Prior Variance
$\mu^{(k=0)}$	-	-	0	0	1	0
$\sigma^{(k=0)}$	-	-	Log (0.4)	1	Log (0.1)	1
φ	-	-	- ∞	0	Logit (0.1)	2
m	-	-	0	0	1	0
κ	Log (1)	0	Log (1)	0	-	-
ω	-	-	-3	4	-6	4

ST-K1	Prior Mean	Prior Variance
μ	Log (1)	0.5
$\hat{r}$	Log (1)	0
$\hat{v}$	Logit (0.5)	0
h	Logit (0.005)	1

RW	Prior Mean	Prior Variance
$v^{(k=0)}$	Logit (0.5)	0
α	Logit (0.5)	1

Response Model	Prior Mean	Prior Variance

	Level 1		Level 2		Level 3	
HGF	Prior Mean	Prior Variance	Prior Mean	Prior Variance	Prior Mean	Prior Variance
ζ	Log (1)	16				
β	Log (16)	16				

Table A. 3. Mean posterior estimates of parameters of interest estimated from winning model.

	ω_{2card}	ω_{2gaze}	ω_{3card}	ω_{3gaze}	β	ζ
Mean	-3.317	-3.193	-5.889	-6.000	1.402	-1.343
Std. Deviation	1.765	1.931	0.3263	0.09111	0.7764	2.609
Minimum	-9.166	-8.899	-6.105	-6.322	-1.551	-4.574
Maximum	-0.5981	-0.6706	-4.340	-5.702	3.393	3.721
Missing values	0	0	0	0	0	0

2. Results

Table A. 4. fMRI results for differential responses to $\hat{\mu}_{1.gaze}^{(t)}$as a function of the computational weighting parameter ζ.

		Cluster		MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
White Matter	0.03	268	4.32	36	-76	-4
R Inferior Occipital Gyrus			4.21	46	-82	-2

Table A. 5. fMRI results for predicted accuracy of advice ($\hat{\mu}_{1.gaze}^{(t)}$) during the choice phase when estimated in GLM containing all parametric modulators.

		Cluster		MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
R Inferior Temporal Gyrus	0	1553	7.29	54	-60	-6
R Fusiform Gyrus			3.93	42	-72	-18
R Cerebelum (VI)			3.7	46	-34	-32
L SupraMarginal Gyrus	0	717	5.75	-56	-28	34
L Inferior Parietal Lobule			4.13	-54	-36	50
L Inferior Temporal Gyrus	0.003	450	5.72	-50	-68	-8
R SupraMarginal Gyrus	0	958	5.24	54	-20	28
R Postcentral Gyrus			5.22	64	-14	26
R Superior Temporal Gyrus			3.8	60	-32	22
L Inferior Parietal Lobule	0	2448	5.19	-32	-42	52
L Paracentral Lobule			5.11	-18	-14	66
L Precentral Gyrus			5.1	-44	-10	56
L Postcentral Gyrus			5.05	-34	-10	58

L Posterior-Medial Frontal			4.8	-2	-6	68
L Superior Parietal Lobule			4.57	-28	-44	62
R Precentral Gyrus	0.017	318	4.92	58	10	30
R Inferior Frontal Gyrus (p. Opercularis)			3.93	52	8	18
RPrecentral Gyrus	0.001	567	4.51	20	-18	70
R Middle Frontal Gyrus			4.27	40	-6	60
R Superior Frontal Gyrus			4.09	26	-6	66
R Caudate Nucleus	0.005	410	4.44	12	8	-12
R Putamen			4.17	18	16	-10
R Pallidum			3.99	22	4	0
R Superior Orbital Gyrus			3.78	18	22	-18
R Olfactory cortex			3.46	10	18	-12

Table A. 6. Neural correlates of differential responses to $\hat{\mu}_{1.gaze}^{(t)}$as a function of the subjective report of having used the gaze and as a function of computational weighting parameter ζ (in GLM containing all parametric modulators).

		Cluster		MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
Subjective Report						
R Rectal Gyrus	0.001	557	4.75	20	16	-12
R Putamen			4.66	32	10	2
R Insula Lobe			3.91	38	6	12
R Rolandic Operculum			3.87	46	2	12
White Matter	0.033	262	4.52	10	-56	-30
R Cerebelum (IX)			4.49	14	-46	-38
R Cerebelum (VIII)			4.11	20	-54	-42
White Matter	0.04	248	4.21	26	-12	-8
R Putamen			4	34	-12	-6
R Hippocampus			3.97	28	-22	-14
R Pallidum			3.95	18	-4	-2
R Thalamus			3.37	16	-18	-2
Computational Weighting Parameter ζ						
R Inferior Occipital Gyrus	0.013	330	4.55	36	-74	-6
L Postcentral Gyrus	0.006	391	4.34	-18	-32	68
L Precentral Gyrus			4.01	-28	-30	60
L Paracentral Lobule			3.94	-14	-32	82

Table A. 7. FMRI results on negative contrast of $|\delta_{1card}^{(t)}|$ during outcome phases where advice was incorrect and during all outcomes.

		Cluster		MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
All outcomes						
R Caudate Nucleus	0	813	6.28	12	16	-8
R Amygdala			4.89	26	0	-14

R ParaHippocampal Gyrus			4.32	22	-4	-24
R Putamen			3.84	32	-2	8
L Middle Occipital Gyrus	0	1160	5.18	-32	-90	18
L Precuneus			4.84	-8	-58	14
L Middle Occipital Gyrus			4.3	-26	-80	34
L Cuneus			4.15	-18	-58	22
L Middle Occipital Gyrus			3.63	-42	-86	24
L Inferior Temporal Gyrus	0	1464	5.04	-36	-42	-16
L Inferior Occipital Gyrus			4.84	-50	-74	-6
L Middle Occipital Gyrus			4.04	-38	-78	-2
L Hippocampus			4.04	-28	-20	-16
L Middle Temporal Gyrus			4.04	-42	-60	-4
R Middle Cingulate Cortex	0.004	348	4.78	4	-40	32
L Posterior Cingulate Cortex			3.87	-6	-46	30
L Putamen	0.025	240	4.53	-20	10	-6
L Olfactory cortex			3.74	-20	4	-14
L Caudate Nucleus			3.56	-10	14	-10
L Hippocampus			3.33	-18	-6	-14
White Matter	0	624	4.45	18	38	-10
L Mid Orbital Gyrus			4.1	-2	50	-8
R Rectal Gyrus			3.96	2	50	-20
L Paracentral Lobule	0.045	206	4.05	-2	-26	52
R Posterior-Medial Frontal			3.84	6	-16	58
L Middle Cingulate Cortex			3.49	0	-26	44
Advice incorrect						
L Inferior Temporal Gyrus	0	2961	6.14	-46	-58	-14
L Fusiform Gyrus			5.97	-38	-50	-16
L Inferior Occipital Gyrus			5.52	-48	-74	-4
L Middle Occipital Gyrus			4.63	-40	-76	-2
L Superior Occipital Gyrus			4.6	-26	-72	22
L Lingual Gyrus			4.19	-26	-96	-16
R Amygdala	0.002	402	5.36	18	4	-16
R Rectal Gyrus			3.75	12	14	-14
R Putamen			3.7	18	18	-8
L Olfactory cortex	0.032	230	4.97	-22	4	-14
L Putamen			4.27	-24	10	0
R Inferior Occipital Gyrus	0	1017	4.75	30	-100	-4
R Middle Occipital Gyrus			4.74	32	-86	20
R Middle Temporal Gyrus			4.5	46	-48	-4
R Inferior Temporal Gyrus			3.96	52	-58	-8
R Inferior Frontal Gyrus (p. Opercularis)	0.006	336	4.7	52	10	26
RPrecentral Gyrus			4.2	62	6	32
R Mid Orbital Gyrus	0.029	234	4.22	4	48	-6
R Superior Medial Gyrus			3.89	4	58	0
L Middle Cigulate Cortex	0.008	315	4.16	-2	-38	34

Neural correlates of face fixations during choice

Increased fixations on the face during choice correlated with increased activity in the right superior temporal gyrus, inferior parietal lobule and angular gyrus (Fig A.1. Table A.5).

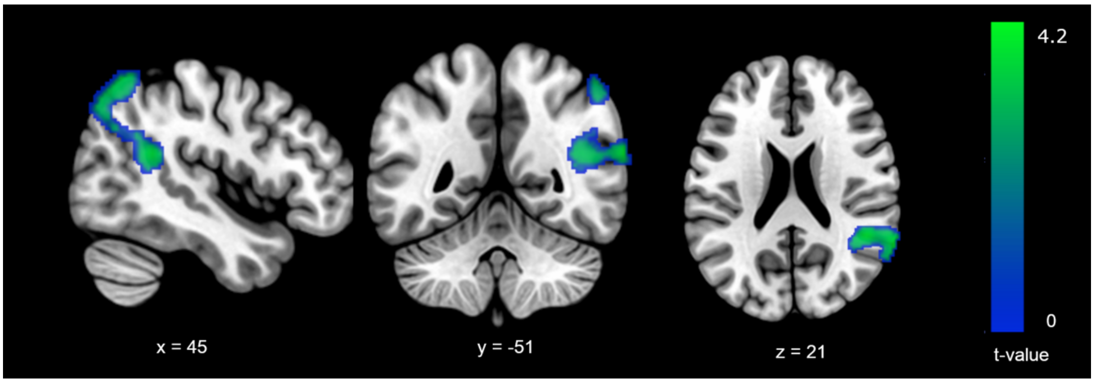


Fig. A. 1. Neural correlates of face fixations during choice. The positive contrast on the parametric modulator reflects BOLD activity in regions correlated with proportion of gaze fixations on the social cue. Cluster-forming threshold: $p<0.005$. cluster level threshold $p< 0.05$. FWE corrected. [x y z] coordinates refer to the MNI coordinates of the respective slices. See Table A. 6 for further information on cluster extents and peak voxel coordinates.

Table A. 5. fMRI results for face fixations during choice.

	Cluster			MNI coordinates		
Region (left/right)	Pcluster	k	Tpeak	x	y	z
R Superior Temporal Gyrus	0.003	945	4.31	48	-46	20
R Inferior Parietal Lobule			3.79	44	-62	58
R Angular Gyrus			3.25	60	-60	36

3 Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder

This chapter includes the second, behavioural study that adopted a computational psychiatry approach to investigate social and non-social learning in decision-making in patients with BPD, SCZ and MDD. The findings demonstrated a commonality in decision-making in patients with BPD and SCZ that was characterised by an excessive reliance on the social information during-decision-making. In addition, the results demonstrate a distinguishing learning pattern in patients with BPD that was characterised by blunted learning of the probabilistic contingencies of both social and non-social outcomes, in conjunction with an exaggerated learning about their volatility. The study highlights the potential for computational modelling in individually estimating aberrant social learning and decision-making patterns and may improve our understanding of the fundamental social impairments in psychiatric disorders. The manuscript has been submitted for publication.

Authors:
Henco L, Diaconescu AO, Lahnakoski JM, Brandi M-L, Hörmann S, Hennings J, Hasan A, Papazova I, Strube W, Bolis D, Schilbach L & Mathys C. 2020. Submitted for publication.

Contributions:
The author of this thesis is the first author of the publication; AOD and LS designed the research; JML programmed the online experiment; **LH**, SH, JH, AH, IP and WS recruited participants and collected data; **LH** performed the computational data analysis with help of CM under the supervision of LS; **LH** wrote the manuscript; all authors reviewed and edited the manuscript. LS was the supervisor of the project and provided funding.

Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder

Lara Henco[1,2*], Andreea O. Diaconescu[3,4,5], Juha M. Lahnakoski[1,6], Marie-Luise Brandi[1], Sophia Hörmann[1], Johannes Hennings[7], Alkomiet Hasan[8], Irina Papazova[8],Wolfgang Strube[8], Dimitris Bolis[1,9], Leonhard Schilbach[1,2,9,10 ¶] & Christoph Mathys[4,11,12 ¶]

[1] Independent Max Planck Research Group for Social Neuroscience, Max Planck Institute of Psychiatry, Munich, Germany

[2] Graduate School for Systemic Neurosciences, Munich, Germany

[3] Department of Psychiatry (UPK), University of Basel, Basel, Switzerland

[4] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland

[5] Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), University of Toronto, Canada

[6] Biomarker Development Research Group, Institute for Neuroscience and Medicine, Brain & Behaviour (INM-7), Forschungszentrum Jülich, Jülich, Germany

[7] kbo-Isar-Amper-Klinikum Munich-East, Munich/Haar, Germany

[8] Department of Psychiatry and Psychotherapy, University Hospital Munich, LMU Munich, Munich, Germany

[9] International Max Planck Research School for Translational Psychiatry (IMPRS-TP), Munich, Germany

[10] Department of Psychiatry and Psychotherapy, University Hospital Düsseldorf, Heinrich Heine-University Düsseldorf, Düsseldorf, Germany

[11] Scuola Internazionale Superiore di Studi Avanzati (SISSA),Trieste, Italy

[12] Interacting Minds Centre, Aarhus University, Aarhus, Denmark

[*] Corresponding author

E-mail: lara_henco@psych.mpg.de

¶ The authors contributed equally to this work

Abstract

Psychiatric disorders are ubiquitously characterized by debilitating impairments in social functioning. These difficulties are thought to emerge from aberrant social inference. In order to elucidate on the computational mechanisms that may underlie such aberrations, patients diagnosed with major depressive disorder (N=29), schizophrenia (N=31), and borderline personality disorder (N=31), and healthy controls (N=34) performed a probabilistic reward learning task in which participants could learn from social and nonsocial information. We applied computational modeling of behavior to assess learning and decision-making parameters estimated for each participant. Participants with borderline personality disorder showed slower learning than healthy controls from both social and non-social information but increased learning of environmental volatility for both types of information. Compared to controls and major depressive disorder patients, borderline personality disorder and schizophrenia patients both showed more reliance on their social, relative to their non-social predictions when making choices. Major depression patients did not differ significantly from controls. This is the first study to apply a computational approach to social and non-social inference transdiagnostically across three different psychiatric patient groups. Computational modeling revealed impaired learning from social and non-social information in borderline personality disorder characterized by an exaggerated sensitivity to changes in environmental volatility. Compared to controls, patients with borderline personality disorder and schizophrenia showed an over-reliance on their beliefs about the predictive value of social relative to non-social information during decision-making. This supports a mechanistic computational account of the exaggerated need to make sense of and rely on other people's minds (over-mentalizing) which is prominent in both disorders.

Author Summary

People suffering from psychiatric disorders frequently experience difficulties in social interaction, such as an impaired ability to use social signals to build representations of others and use these to guide behavior. Compuational models of learning and decision-making enable the characterization of individual patterns in learning mechanisms that may be disorder-specific or disorder-general. We employed this approach to investigate the behavior of healthy participants and patients diagnosed with depression, schizophrenia, and borderline personality disorder while they performed a probabilistic reward learning task which included a social component.

We found that learning in patients with borderline personality disorder was characterized by a reduced flexibility in the weighting of newly obtained social and non-social information according to its predictive value. Instead, we found exagerrated learning of the volatility of social and non-social information. Additionally, we found a pattern shared between patients with borderline personality disorder and schizophrenia who both showed an over-reliance on predictions about social relative to non-social information during decision-making. Our modeling provides a computational account of the exaggerated need to make sense of and rely on other people's minds, which is prominent in both disorders.

Introduction

Impairments in social cognition are frequently experienced by people suffering from a psychiatric disorder. For instance, patients with major depressive disorder (MDD) and schizophrenia (SCZ) show a reduction in (social) reward sensitivity and motivation to engage in social interactions [1–5]. Despite high levels of social anhedonia, patients with SCZ show a tendency to over-interpret the meaning of social signals [6]. Individuals with borderline personality disorder (BPD) suffer from rapidly changing beliefs about others that polarise between approach and rejection [7]. Together, these impairments are associated with aberrant inferences/beliefs about oneself and the social environment.

In computational terms, the emergence of aberrant inference can be ascribed to an impaired ability to adjust learning in response to environmental changes [8]. Bayesian learning models allow for a parsimonious algorithmic description of changes in beliefs relevant for accurate inference: belief updates can be written as a surprise signal (prediction error) weighted by a learning rate [9]. The learning rate depends on the ratio between the precision of the sensory data and the precision of the prior belief [10,11]. Whereas healthy participants increase their learning rate more strongly in volatile compared to stable environments [12,13], patients with autism do so less owing to an over-estimation of environmental volatility [8]. Impairments in the estimation of environmental volatility have also been proposed as a mechanism for psychosis and SCZ [14,15] as well as MDD [16]. One recent study found that, unlike healthy controls, participants with BPD did not show an increase in learning when social and reward contingencies became volatile [17]. The authors suggested that this might be due to higher expected baseline volatility in participants with BPD. However, the computational model employed in that study did not explicitly model beliefs about volatility.

Adopting previous suggestions of aberrant volatility learning in psychiatric disorders and its role in impaired probability learning, the current study employed Bayesian hierarchical modeling to investigate probabilistic (social) inference in a volatile context learning across three major psychiatric disorders, which have previously been associated with social dysfunction: MDD, SCZ and BPD. Here, the current study investigated whether volatility and probability learning is equally affected when inferring on the hidden states of non-social and social outcomes across the three different disorders. We further asked whether aberrant social learning and decision-making were associated with differences in social anhedonia.

To this end, we adopted a probabilistic reward learning task (introduced in [18]), in which participants could learn from two types of information: non-social and social information. In order to probe the spontaneous rather than explicitly instructed use of social information as in previous social learning studies [12,13,19,20], we did not explicitly tell participants to learn about the social information. We used the hierarchical Gaussian filter (HGF; [10,11]) to obtain a profile of each participant's particular way of updating beliefs when receiving social and non-social information while making decisions in a volatile context. The HGF is a generic hierarchical Bayesian inference model for volatile environments with parameters that reflect individual variations in cognitive style. We went beyond other recent computational psychiatry studies using the HGF (e.g., [8,21–25]) in that we used two parallel HGF hierarchies for social and non-social aspects of the environment (cf. [26,27]). Our modeling framework was specifically designed also to quantify the relative weight participants accorded their beliefs about the predictive value of social compared to non-social information in decision-making.

Materials and Methods

Participants

Patients were recruited for the present study after an independent and experienced clinician diagnosed them using ICD-10 criteria for 1) a depressive episode (F32), schizophrenia (F20) and emotionally unstable personality disorder (F60.3). HC and patients with MDD were recruited through the Max Planck Institute of Psychiatry. Patients with SCZ were recruited at the Department of Psychiatry and Psychotherapy at the University Hospital Munich. Patients with BPD were recruited at the kbo-Isar-Amper-Klinikum in Haar, Munich. All participants were naïve to the purpose of the experiment and provided informed consent to take part in the study after a written and verbal explanation of the study procedure. The study was in line with the Declaration of Helsinki and approval for the experimental protocol was granted by the local ethics committee of the Medical Faculty of the Ludwig-Maximilians University of Munich. Detailed exclusion criteria are listed in the Supplementary Methods. Participants were chosen prior to analysis such that groups were matched for age (χ^2=5.302, P=0.151; Kruskal-Wallis one-way non-parametric ANOVA because of difference in age variance between groups, see Table S1). Exclusion criteria were a history of neurological disease or injury, reported substance abuse at the time of the investigation, a history of electroconvulsive therapy, and diagnoses of comorbid personality disorder in the case of MDD and SCZ. Furthermore, 9 participants had to be excluded from the analysis due to one of the following reasons: unsaved data due to technical problems (1 HC, 2 BPD). Prior participation in another study which involved the same paradigm (1 HC), always picking the card with the higher reward value (1 HC), either following (1 SCZ) or going against (1 BPD) the gaze on more than 95% of trials (indicating a learning-free strategy), interruption of the task (1 SCZ), change to the diagnosis following study participation (1 MDD). The final sample consisted of 31 HC, 28 MDD, 29 SCZ and 28 BPD. We additionally acquired psychometric data (Supplementary Methods and Table S1) to further characterize the participants: All patients were asked to fill out questionnaires measuring autistic traits with the autism spectrum quotient (AQ [28]) and social anhedonia symptoms with the Anticipatory and Consummatory Interpersonal Pleasure Scale (ACIPS; [29]). We additionally assessed positive and negative symptoms using the Positive and Negative Syndrome Scale (PANSS [30]) and mood symptoms using the Calgary Depression Scale for Schizophrenia (CDSS [31]) in patients with SCZ. To assess the severity of Borderline Personality Disorder we used the short version of the Borderline Symptom List (BSL-23 [32]). Additional questionnaires were employed but analyzed within the scope of a different study and therefore not presented here. Demographic data as well as details regarding the medication can be seen in the Supplementary Methods, S2 Table).

Experimental paradigm and procedure

After giving informed consent, participants were seated in front of a computer screen in a quiet room where they received the task instructions. In the probabilistic learning task introduced in [26], participants were asked to choose between one of two cards (blue or green) in order to maximize their score which was converted into a monetary reward (1-6 €) that was added to participants' compensation at the end of the task. An animated face was displayed between the cards, which first gazed down, then up towards the participant, before it shifted its gaze towards one of the cards (Fig 1A). The blue and green card appeared randomly on the left and right side from the face and participants responded using 'a' or 'l' on a German QWERTZ keyboard. When a response was logged within the allowed time (6000 ms), the chosen card was marked for 1000 ms until the outcome (correct: green check mark/wrong: red cross) was displayed for 1000 ms. When the correct card was chosen, the reward value (1-9) displayed on the card was added to the score. Participants were instructed that these values were not associated with the cards' winning probabilities, but that they might want to choose the card with the higher value if they were completely uncertain about the outcome. When the wrong card was chosen or participants failed to choose a card in the allotted time, the score remained unchanged. Participants were told that the cards had winning probabilities that changed in the course of the experiment but they were not informed about the systematic association between the face animation's gaze and the trial outcome. Specifically, they were not told that the probability with which the face animation pointed towards the winning card on a given trial varied systematically throughout the task according to the schedule given in Fig 1B. Instead, we simply told participants that the face was integrated into the task to make it more interesting. The probabilistic schedules for social and non-social information were independent from each other in order to estimate participant-specific learning rates separately for both types of information. In the first half of the experiment (trials 1–60), the card winning probabilities were stable, whereas in the second half (trials 61–120) they changed (volatile phase). The social cue had a stable contingency during trials 1-30 and trials 71-120, whereas contingency was volatile during trials 31-70. We used two types of schedules for the social cue which were each presented to half of the participants. In one schedule (depicted in Fig 1B), the probability of the social cue looking towards the winning card was 73% in the first stable phase (trial 1-30) and therefore started as congruent to the winning card (congruent-first). The second probability schedule was flipped, so that the probability of the social cue looking towards the winning card was 27% in the first stable phase (incongruent-first). In total, 15 control participants received the congruent-first schedule, 15 participants with MDD, 14 with SCZ and 15 with BPD. Positions of the cards on the screen (blue left or right) were determined randomly. The task was programmed and presented with PsyToolkit [33].

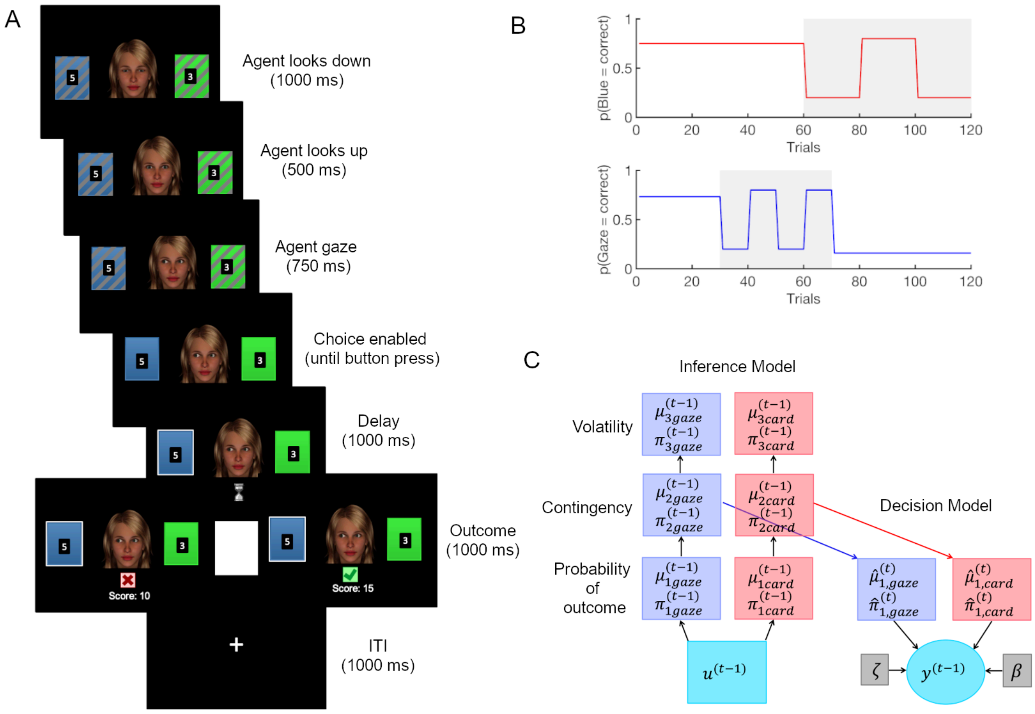

A
Agent looks down
(1000 ms)
Agent looks up
(500 ms)
Agent gaze
(750 ms)
Choice enabled
(until button press)
Delay
(1000 ms)
Outcome
(1000 ms)
Score: 10
Score: 15
ITI
(1000 ms)
5
3
B
p(Blue = correct)
p(Gaze = correct)
1
0.5
0
0
20
40
60
80
100
120
Trials
C
Inference Model
Volatility
Contingency
Probability of outcome
Decision Model
$\mu_{3gaze}^{(t-1)}$ $\pi_{3gaze}^{(t-1)}$
$\mu_{3card}^{(t-1)}$ $\pi_{3card}^{(t-1)}$
$\mu_{2gaze}^{(t-1)}$ $\pi_{2gaze}^{(t-1)}$
$\mu_{2card}^{(t-1)}$ $\pi_{2card}^{(t-1)}$
$\mu_{1gaze}^{(t-1)}$ $\pi_{1gaze}^{(t-1)}$
$\mu_{1card}^{(t-1)}$ $\pi_{1card}^{(t-1)}$
$\hat{\mu}_{1,gaze}^{(t)}$ $\hat{\pi}_{1,gaze}^{(t)}$
$\hat{\mu}_{1,card}^{(t)}$ $\hat{\pi}_{1,card}^{(t)}$
$u^{(t-1)}$
ζ
$y^{(t-1)}$
β

Fig 1. Task design and computational decision and inference model. (A) Participants were asked to make a choice between blue and green cards after the gray shading on the colored rectangles (cards) had disappeared (i.e., 750

ms after the face shifted its gaze towards one of the cards). After a delay phase, the outcome was presented (correct/wrong). If the choice was correct, the reward amount (number on the chosen card) was added to a cumulative score. The task consisted of 120 trials. (B) Probability schedules from which outcomes were drawn. Volatile phases are marked in grey. (C) Posteriors are deterministic functions of predictions and outcomes. Predictions in turn are deterministic functions of the posteriors of previous trials. Decisions $y^{(t)}$ are probabilistically determined by predictions and decision model parameters ζ and β. Deterministic quantities are presented as boxes and probabilistic quantities in circles.

Computational Modeling

Observing the observer

We modeled behavior in the 'observing the observer' (OTO) framework [34,35]. This entails a *response model*, which probabilistically predicts a participant's choices based on his/her inferred beliefs, and a *perceptual model*, on which the response model depends because it describes the trajectories of participants' inferred beliefs based on experimental inputs. The OTO framework is conceptually very similar to the idea of *inverse reinforcement learning* [36].

Perceptual Model

As a perceptual model we used two parallel HGF hierarchies to represent concurrent hierarchical learning about the social (predictive value of gaze) and non-social (predictive value of card color) aspects of the task environment. The HGF is an inference model resulting from the inversion of a generative model in which states of the world are coupled in a three-level hierarchy: At the lowest level of the generative model, ${x_1}_{gaze}$ and ${x_1}_{card}$ represent the two inputs in a binary form (social cue: 1=correct, 0=incorrect; card outcome: 1=blue wins, 0=green wins). Level ${x_2}_{gaze}$ and ${x_2}_{card}$ represent the tendency of the gaze to be correct and the tendency of the blue card to win. State ${x_2}_{gaze}$ and ${x_2}_{card}$ evolve as first-order autoregressive (AR(1)) processes with a step size determined by the state at the third level. Level ${x_3}_{gaze}$ and ${x_3}_{card}$ represent the log-volatility of the two tendencies and also evolve as first-order autoregressive (AR(1)) processes. The probabilities of ${x_1}_{gaze} = 1$ and ${x_1}_{card} = 1$ are the logistic sigmoid transformations of ${x_2}_{gaze}$ and ${x_2}_{card}$ (Equation 1).

$$p\left({x_1}^{(t)} = 1\right) = \frac{1}{1+\exp\left(-{x_2}^{(t)}\right)} \qquad (1)$$

Participants' responses y were coded with respect to the congruency with the 'advice' (1=follow; 0=not follow) and were used to invert the model in order to infer the belief trajectories at all three levels $i = 1,2,3$.

On every trial k, the beliefs $\mu_i^{(k)}$ (and their precisions $\pi_i^{(k)}$) about the environmental states at the i-th level are updated via prediction errors $\delta_{i-1}^{(k)}$ from the level below weighted by a precision ratio $\psi_i^{(k)}$(Equation2-3). This means that belief updates are larger (due to higher precision weights) when the precision of the posterior belief (${\pi_2}^{(k)}$ or $\pi_3^{(k)}$) is low and the precision of the prediction $\hat{\pi}_2^{(k)}$is high. Consequently, prediction errors are weighted more during phases of high volatility (cf. Figure S1, panel C, dotted blue trajectory). For the analysis, we used $q(\psi_2^{(k)})$ (Equation 4), which is a transformation of $\psi_2^{(k)}$ (Equation 3) (cf. [37], supplementary material) that corrects for the sigmoid mapping between first and second level, effectively making $q(\psi_2^{(k)})$ an uncertainty (inverse precision) measure for first-level beliefs.

$$\Delta\mu_i^{(k)} \propto \psi_i^{(k)}\delta_{i-1}^{(k)} \quad (i = 2,3) \qquad (2)$$

$$\psi_2^{(k)} = \frac{1}{{\pi_2}^{(k)}} \qquad (3)$$

$$q(\psi_2^{(k)}) = \psi_2^{(k)} \; (s\left(\mu_2^{(k)}\right)\left(1 - s\left(\mu_2^{(k)}\right)\right)) \qquad (4)$$

$$\psi_3^{(k)} = \frac{\hat{\pi}_2^{(k)}}{\pi_3^{(k)}} \qquad (5)$$

Participant-specific parameters ω_{2card} and ω_{2gaze} represent the learning rates at the second level, i.e. the speed at which association strengths change. Correspondingly, ω_{3card} and ω_{3gaze} represent the learning rates of the volatilities.

Response Model

In the response model a combined belief $b^{(t)}$ (Equation 6) was mapped onto decisions, which resulted from a combination of both the inferred prediction $\hat{\mu}_{1,gaze}^{(t)}$ that the face animation's gaze will go to the winning card and the inferred prediction $\hat{\mu}_{1,card}^{(t)}$ that the color of the card that the gaze went to would win (see example in S1 Fig). The inferred prediction $\hat{\mu}_{1,gaze}^{(t)}$ and

$\hat{\mu}_{1,card}^{(t)}$ were weighted by $w_{gaze}^{(t)}$ and $w_{card}^{(t)}$ (Equation 7-8), which are functions of the respective precisions ($\hat{\pi}_{1,gaze}^{(t)}$ and $\hat{\pi}_{1,card}^{(t)}$, Equation 9-10). The precisions (Equation 9-10) represent the inverse variances a Bernoulli distribution of $\hat{\mu}_{1,gaze}^{(t)}$ and $\hat{\mu}_{1,card}^{(t)}$.

The constant parameter ζ represents the weight on the precision of the social prediction compared to the precision of the non-social prediction (Equation 7). In other words, this parameter describes the propensity to weight the social over the non-social information. We investigated the effect of varying the social weighting factor ζ, by simulating the combined belief $b^{(t)}$ (Equation 6) of agents with same perceptual parameters (fixed at prior values as depicted in S3 Table) but different ζ values (Fig 5B&C).

$$b^{(t)} = w_{gaze}^{(t)}\,\hat{\mu}_{1,gaze}^{(t)} + w_{card}^{(t)}\,\hat{\mu}_{1,card}^{(t)} \qquad (6)$$

$$w_{gaze}^{(t)} = \frac{\zeta\hat{\pi}_{1,gaze}^{(t)}}{\zeta\hat{\pi}_{1,gaze}^{(t)}+\hat{\pi}_{1,card}^{(t)}} \qquad (7)$$

$$w_{card}^{(t)} = \frac{\hat{\pi}_{1,card}^{(t)}}{\zeta\hat{\pi}_{1,gaze}^{(t)}+\hat{\pi}_{1,card}^{(t)}} \qquad (8)$$

$$\hat{\pi}_{1,gaze}^{(t)} = \frac{1}{\hat{\mu}_{1,gaze}^{(t)}\,(1-\hat{\mu}_{1,gaze}^{(t)})} \qquad (9)$$

$$\hat{\pi}_{1,card}^{(t)} = \frac{1}{\hat{\mu}_{1,card}^{(t)}\,(1-\hat{\mu}_{1,card}^{(t)})} \qquad (10)$$

In the response model, used the combined belief $b^{(t)}$ (Equation 6) in a logistic sigmoid (softmax) function to model the probability $Prob_{gaze}^{(t)}$ (Equation 11). In this function, the belief was weighted by the predicted reward of the card when the advice is taken r_{gaze} or not $r_{notgaze}$ (Equation 11).

$$prob_{gaze} = p\left(y^{(t)} = 1\right) = 1/\left(1 + \exp\left(-\gamma^{(t)}\left(r_{gaze}^{(t)}b^{(t)} - r_{notgaze}^{(t)}\left(1 - b^{(t)}\right)\right)\right)\right) \qquad (11)$$

The mapping of beliefs onto actions varied as a function of the inverse decision temperature $\gamma^{(t)}$, where large $\gamma^{(t)}$ implied a high alignment between belief and choice (low decision noise) and a smaller $\gamma^{(t)}$ a low alignment between belief and choice (high decision noise). Our four different response models varied in terms of how $\gamma^{(t)}$ was defined. In response model 1, $\gamma^{(t)}$ was a combination of the log-volatility of the third level for both cues combined with constant participant-specific decision noise β (Equation 12). In response model 2, $\gamma^{(t)}$ was a combination of the log-volatility of the third level for the social cue and participant-specific decision noise (Equation 13) and in response model 3 $\gamma^{(t)}$ was a combination of the log-volatility of the third level for the non-social cue and participant-specific decision noise (Equation 14). In model 4, $\gamma^{(t)}$ only included the participant-specific decision noise (Equation 15).

1) $\gamma^{(t)} = \beta \exp\left(-\hat{\mu}_{3,card}^{(t)} - \hat{\mu}_{3,gaze}^{(t)}\right)$ (12)

2) $\gamma^{(t)} = \beta \exp\left(-\hat{\mu}_{3,gaze}^{(t)}\right)$ (13)

3) $\gamma^{(t)} = \beta \exp\left(-\hat{\mu}_{3,card}^{(t)}\right)$ (14)

4)$\gamma^{(t)} = \beta$ (15)

We used the HGF toolbox, version 4.1, which is part of the software package TAPAS (https://translationalneuromodeling.github.io/tapas) for parameter estimation. We fitted six alternate combinations of perceptual and response models, which were subjected to random-effects Bayesian Model Selection [38,39] (spm_BMS in SPM12; http://www.fil.ion.ucl.uk/spm), cf. Supplementary Methods. The HGF was compared against a Sutton K1 model [40] and a Rescorla Wagner learning model with a fixed learning rate [41]. We used the implementation of these models in the HGF toolbox and adjusted them in the same way as the HGF so that the model engages concurrent parallel learning about the social and non-social aspects of the task environment. These widespread and powerful perceptual models were chosen for comparison as in previous studies [24,42]. The HGF was combined with all four response models. The non-hierarchical models were combined with response model 4 only, owing to the lack of third-level belief trajectories. Details of the prior settings of all models can be seen in S3 Table.

Model comparison and validity

The log model evidence (LME) for each participant and each model were subjected to Bayesian Model Selection [38,39] (spm_BMS in SPM12) in order to estimate the expected posterior probabilities (EXP_P), the exceedance probability (XP) and the protected exceedance probability (PXP).

Posterior predictive validity of model parameters and task performance

To test the robustness of the model, we simulated responses based on the estimated parameters from the best fitting model and checked whether these responses produce the same group differences that were observed in the real behavioral data, i.e. different performance accuracy. For this analysis, behavioral responses were simulated based on the posterior estimates of 60 participants (we simulated 15 randomly sampled participants of each of the four groups 10 times). These simulated responses were used to calculate performance (i.e., % of accurate responses) which was then entered into a one-way ANOVA as described in the methods (statistical analysis).

Statistical Analysis

Performance (% correct responses) was subjected to a one-way ANOVA with group (HC vs. MDD vs. SCZ vs. BPD) and schedule (congruent first vs. incongruent first) as between-subject factors. Advice taking (advice followed or not on a given trial) was subjected to a mixed ANOVA with social accuracy (high vs. low) and schedule stability (stable vs. volatile) as within-subject factors. Group (HC vs. MDD vs. SCZ vs. BPD) and schedule (congruent first vs. incongruent first) were included as between-subject factors.

Mean precision weights on the second and third level ($q(\psi_2)$ and ψ_3) separately entered two mixed ANOVAs as dependent variables with schedule stability as a within-subject factor (stable vs.volatile), information type as within participants factor (social vs. non-social). The group (HC vs. MDD vs. SCZ vs. BPD) and schedule (congruent first vs. incongruent first) were between subject factors.

We subjected the posterior estimate for ζ to a one-way ANOVA with group (HC vs. MDD vs. SCZ vs. BPD) as between-subject factor and schedule (congruent first vs. incongruent first) as a covariate.

We hypothesized that social anhedonia (measured by the Anticipatory and Consummatory Interpersonal Pleasure Scale, ACIPS) would be associated with a reduction in learning in the social domain. To test this, we first performed a one-way ANOVA with ACIPS scores as dependent variable and group as the factor (HC vs. MDD vs. SCZ vs. BPD) followed by a multivariate regression with ACIPS as dependent variable and the social learning rates ω_{2gaze} and the weighting factor ζ as predictors of social learning and decision making. The group factor (HC vs. MDD vs. SCZ vs. BPD) was entered as covariate. This analysis was done for all participants who completed the ACIPS questionnaire (n=106 of n_{total}=116).

All ANOVA post hoc *t* tests were Bonferroni-corrected for multiple comparisons. All *p*-values are two-tailed with a significance threshold of *p* <.05. Statistical tests were performed using JASP (Version 0.9 2.0; https://jasp-stats.org/) or Matlab (Version 2018b; https://mathworks.com).

Results

Behavior

There was a significant difference between the groups in the overall performance, i.e. % of correct responses (*F*(3,108)=7.504, *p*<0.001, Fig 2A): Post-hoc comparisons showed that both patients with SCZ and BPD performed significantly worse compared to HC and patients with MDD (SCZ–HC *t*=3.781, p_{bonf}=0.002, SCZ–MDD *t*=2.817, p_{bonf}=0.035, BPD–HC *t*=3.732, p_{bonf}=0.002, BPD–MDD *t*=2.78, p_{bonf}=0.038). There was no significant difference in performance between patients with BPD and SCZ (*t*=-0.01, p_{bonf}=1.000) nor between HC and patients with MDD (*t*=0.88, p_{bonf}=1.000). Performance was not significantly affected by the schedule order (congruent first vs. incongruent first; *F*(1,108)=0.027, *p*=0.870) or its interaction with the patient groups (*F*(3,108)=1.302, *p*=0.278).

We found a main effect of social accuracy (*F*(1,108)=227.935, *p*<0.001) whereby participants followed the gaze more during phases of high accuracy compared to phases of low accuracy (*t*=14.94, p_{bonf}<0.001) (Fig 2C). Advice taking was not significantly affected by the schedule stability (*F*(1,108)=0.503, *p*=0.480), indicating that advice taking did not differ between stable and volatile phases. Advice taking was not significantly affected by an interaction between accuracy of the social information and Group (*F*(3,108)=2.222, *p*=0.09), or by an interaction between social accuracy, schedule stability and Group (*F*(3,108)=1.47, *p*=0.227). However, there appeared to be a group difference during the volatile low accuracy trials (rightmost data points in Fig 2.C).

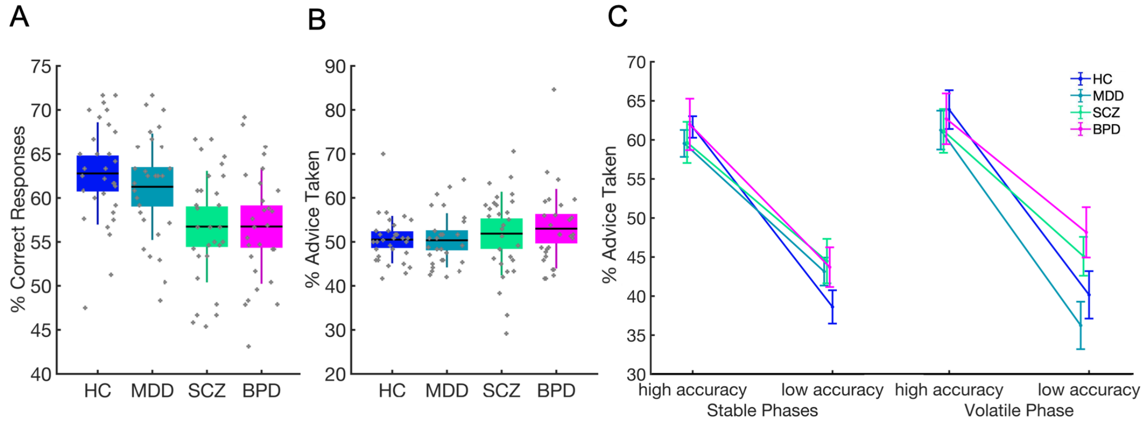


Fig 2. Behavioral results. (A) Patients with SCZ and BPD show poorer performance in task compared to HC and patients with MDD. (B) There was no difference between groups with respect to the percentage of trials in which the advice was taken. Boxes mark 95% confidence intervals and vertical lines standard deviations. (C) Patients with BPD followed the advice significantly more compared to patients with MDD during volatile phases of low accuracy.

Bayesian Model Selection & Validity

Model comparison showed that the HGF including subject specific decision noise as well as the volatility estimate $\hat{\mu}_{3,gaze}$ and $\hat{\mu}_{3,card}$ outperformed the other HGF models as well as the Rescorla Wagner and Sutton-K1 models with subject specific decision noise only (PXP=0.958; XP=0.998). See Table 1 for further details and S4 Table for mean posterior parameter estimates.

Table 1. Bayesian Model Selection results.

BMS	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
EXP_R	0.492	0.067	0.032	0.015	0.028	0.267
XP	0.998	0	0	0	0	0.002
PXP	0.958	0.008	0.008	0.008	0.008	0.01

Posterior model probabilities (EXP_R), Exceedance Probabilities (XP) and Protected Exceedance Probabilities (PXP). Model 1 refers to the HGF combined with response model 1, Model 2 refers to the HGF combined with response model 2, Model 3 refers to the HGF combined with response model 3, Model 4 refers to the HGF combined with response model 4, Model 5 refers to the Sutton K-1 Model combined with response model 4, Model 6 refers to the Rescorla Wagner Model combined with response model 4.

Posterior predictive validity of model parameters and task performance

We simulated responses using the posterior mean parameter values of 60 randomly chosen participants from the best fitting model to demonstrate that this model was capable of reproducing the group differences that were observed in the real behavioral data, i.e. different performance accuracy.

The ANOVA of the simulated data showed that for performance accuracy, as in the real data, there was a main effect of group (*F*(3,592) = 35.776, p<0.001) with significantly lower performance for SCZ and BPD patients compared to HC and patients with MDD (all comparisons: p_{bonf} <.001). As in the real data, performance did not significantly differ between HC and MDD or between BPD and SCZ (HC–MDD p_{bonf} =0.216; BPD–SCZ p_{bonf} =1.000). Whereas there was no significant main effect of Schedule as in the real data, there was a significant Group × Schedule effect (p_{bonf} <.001) which was not observed in the real data. While the real data do point towards an interaction between Group and Schedule, this interaction only reached significance in the simulated data. This is most likely due to an increase of power since every participant was simulated 10 times reaching n=600, compared to n=116 as in the real analysis. For the same reason, the group differences in simulated performance have larger effect sizes compared to the group differences in real performance.

Dynamic learning rates – second level

For the averaged precision weights (i.e., dynamic learning rates) for learning about the social $q(\psi_{2gaze})$ and non-social $q(\psi_{2card})$, we found a main effect of task phase (*F*(1,108)=18.628, *p*<0.001), showing that $q(\psi_2)$ is higher in volatile compared to the stable phases (*t*=-4.419, p_{bonf} <0.001) (Fig 3A). There was no significant interaction between Phase and Information Type, which indicates that $q(\psi_2)$ increases similarly during social and non-social volatility (*F*(1,108)= 1.654, *p*=0.201). However, the increase in $q(\psi_2)$ in volatile phases was stronger when participants received the congruent-first schedule (*F*(1,108)=3.988, *p*=0.048). There was a significant main effect of group (*F*(3,108)=3.939, *p*=0.01), and the post-hoc *t*-tests revealed that participants with BPD showed significantly lower precision weights on the second level compared to HC (*t*=3.346, p_{bonf}=0.007). The difference in $q(\psi_2)$ between groups was not affected by Information type (*F*(3,108)= 0.946, *p*=0.421) or its interaction with Phase (*F*(3,108)= 1.644, *p*=0.184) (for full table of results see S5 Table).

Dynamic learning rates – third level

We found a main effect of task phase on precision weights at the third level ($F(1,108)=125.99$, $p<0.001$), showing that ψ_3 is higher in volatile compared to the stable phases ($t=-10.06$, $p_{\mathrm{bonf}}<0.001$) (Fig 3B). There was a significant main effect of group ($F(3,108)=7.159$, $p<0.001$), and post-hoc t tests showed that participants with BPD showed significantly higher precision weights at the third level compared to all other groups (BPD–HC $t=-4.332$, $p_{\mathrm{bonf}}<.001$; BPD–MDD $t=-3.51$, $p_{\mathrm{bonf}}=.004$; BPD–SCZ $t=-3.21$, $p_{\mathrm{bonf}}=.01$). In addition, there was a significant Phase × Group interaction ($F(3,108) = 6.98$, $p<0.001$) showing that participants with BPD increase their precision weights for both modalities significantly more compared to the other groups when volatility increases. There was a trend of BPD patients showing stronger increases in ψ_3 in response to social compared to non social volatility ($F(3,108)=2.625$, $p=0.054$). The analysis also revealed that ψ_3 were affected by the order of schedule ($F(1,108)=5.118$, $p=0.026$), with ψ_3 higher for participants receiving the incongruent-first schedule (i.e gaze starts of being highly misleading) compared to the congruent-first schedule (i.e gaze starts of being highly helpful). This effect was not modulated by Group ($F(3,108)=2.53$, $p=0.061$). See S6 Table for full table of results.

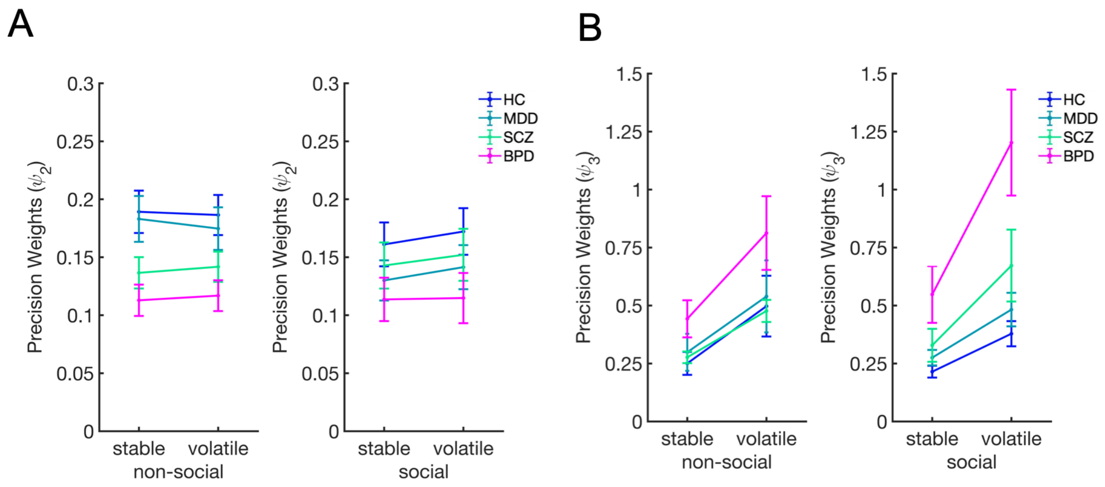


Fig 3. Results for mixed ANOVA using precision weights for updating beliefs about social and non-social contingency and volatility. (A) Precision weights $q(\psi_2)$ and (B) precision weights ψ_3. Overall, $q(\psi_2)$ and ψ_3 increase when transitioning from stable to volatile phase. Patients with BPD show reduced overall $q(\psi_2)$. At same time, patients with BPD show higher ψ_3 compared to the other groups and a more pronounced increase in response to volatility. Bars indicate SEM. See also S2 Fig.

Social Weighting

The parameter ζ was a measure of the weight given to the social prediction relative to the learned non-social prediction. Since ζ was restricted to the positive domain, estimate distributions were analyzed log-space, where they were less skewed. We found significant group differences in log(ζ) (F(3,108)=6.79, p>0.001 (Fig 5A). Both patients with BPD and patients with SCZ showed significantly higher ζ estimates compared to controls (BPD: t=-3.681, p_{bonf}=0.002; SCZ: t=-3.243, p_{bonf}=0.009) but only patients with BPD differed significantly from participants with MDD (BPD: t=-3.036,p_{bonf}=0.018; SCZ: t=-2.602, p_{bonf} =0.063). Patients with MDD did not show any significant differences compared to controls (t=-0.566, p_{bonf}=1). There was a significant main effect of schedule (F(1,108)=8.191, p=0.005), showing that participants receiving the congruency-first schedule had higher ζ compared to participants receiving the incongruency-first schedule (t=-2.862, p_{bonf}=0.005). There was no significant interaction between Group and Schedule (F(3,108)=0.820, p<0.485).

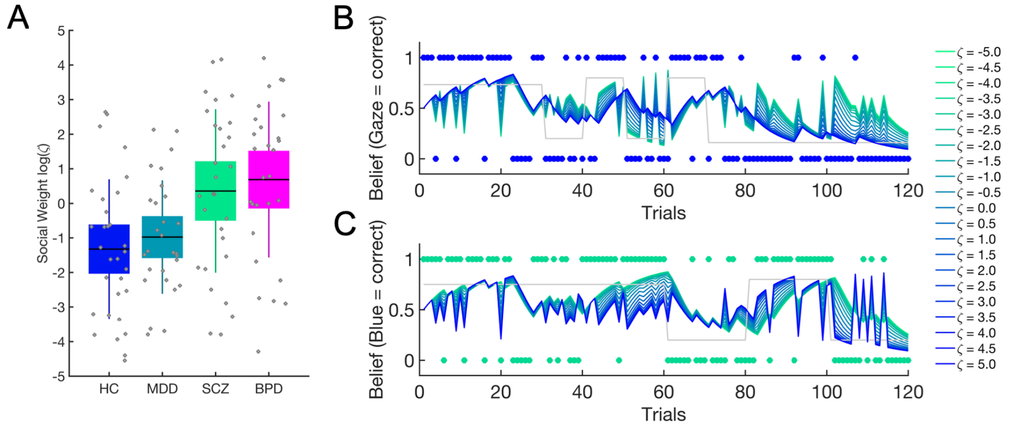


Fig 4. Social weighting factor log(ζ). (A) Patients with BPD gave the social information significantly weight more compared to HC and patients with MDD. Patients with SCZ also had higher ζ compared to HC. Boxes mark 95% confidence intervals and vertical lines standard deviations. B, Simulation results show the impact of varying weighting factor log(ζ) on combined belief $b^{(t)}$ (see methods Equation 1). The combined belief $b^{(t)}$ was simulated for agents with same perceptual parameters but different ζ values (highest values (log(ζ)=5) coded in dark blue, lowest values (log(ζ)=-5) in green). (B) shows that the combined belief $b^{(t)}$ of agents with high ζ values is aligned with the social input structure (blue dots) whereas these agents show a stochastic belief structure with regard to the non-social input structure (green dots) in Panel C. Conversely, agents with low ζ values show a belief structure closely aligned to the non-social input structure (C), and a stochastic belief structure with regard to the social input (Panel B). The grey lines represent the ground truth of the respective probability schedules.

Social Anhedonia

There was a significant difference in the interpersonal pleasure (ACIPS) ratings between the groups (*F*(3,103)=5.719, *p*<.001) (S1 Table). Post hoc *t* tests revealed that HC showed significantly higher ACIPS scores compared to patients with MDD (*t*=3.088, p_{bonf}=.016) and with BPD (*t*=3.802, p_{bonf} =.001) but not with SCZ (*t*=1.833, p_{bonf} =.418). No significant differences were observed between patients with MDD and SCZ (*t*=-1.322, p_{bonf} =1), patients with MDD and BPD (*t*=0.596, p_{bonf}=1), nor patients with SCZ and BPD (*t*=1.978, p_{bonf}=0.304). The multivariate regression using log(ζ) and social learning rate ω_{2gaze} as predictors for ACIPS scores did not show any significant results (R^2=115, *F*(2,106)=0.699, *p*=0.499).

Discussion

This study aimed to improve our understanding of the mechanisms underlying the pervasive interpersonal difficulties in common psychiatric disorders. To achieve this, we used a probabilistic learning task in conjunction with hierarchical Bayesian modeling transdiagnostically in patients with MDD, SCZ, BPD, and healthy controls. The task required participants to perform association learning about non-social contingencies in the presence of a social cue. This allowed us to characterize and quantify the computational aspects of aberrant social inference and decision-making at an individual level. We found that patients with SCZ and BPD showed significantly poorer performance compared to HC and patients with MDD. Patients with MDD performed comparably well to HC. Patients with BPD showed increased precision weighting of prediction errors when learning about the volatility (i.e., the rate of change) in both non-social and social information and a tendency for even higher precision weights when learning about social compared to non-social volatility.

Exaggerated volatility learning in BPD was accompanied by significantly reduced learning rates when simply learning social and non-social contingencies. This accords with a previous finding of blunted social and non-social learning in BPD [43], which was conjectured to result from higher baseline volatility beliefs, causing an impairment at detecting contingency changes needed for accurate inference. Because it was specifically designed to model beliefs about volatility, our modeling approach allowed us to test and nuance this conjecture. Our data indicate that impaired contingency learning in BPD is associated with exaggerated *learning* about environmental volatility instead of a higher baseline volatility belief. A similar pattern has been observed in autism spectrum disorder (ASD) [8]. This commonality may explain the repeated finding of high autism quotient (AQ) values in BPD patients [8], which is confirmed in our sample (cf. S1 Table). Aberrant volatility beliefs in BPD have been suggested to result from unpredictable early relationships [17]. However, this is a less likely explanation in ASD, which points to a different origin of the mechanistic overlap between our findings and those of the aforementioned study [8]. Additionally, this is the second study demonstrating that aberrant learning in BPD not only concerns social, but also non-social information (cf. [43]), which could point to domain-independent learning impairments.

Unlike previous studies on reward [44–46] or volatility [14] learning in SCZ, we did not find significant differences between SCZ and HC in that regard. However, we found that SCZ and BPD patients both weighted their social-domain predictions more strongly when making decisions than HC and MDD. One possible explanation for the lower performance of BPD and SCZ patients is that their stronger reliance on social cues compared to HC and MDD patients is detrimental during the volatile gaze phase, where the reliability of the gaze information is reduced compared to the non-social one.

Our finding that learning in SCZ is not significantly different from learning in HC is in line with a previous study [47] showing intact reward learning but altered weighting of response options in SCZ patients. The computational commonality between SCZ and BPD of over-weighting social-domain predictions, is intriguing because it suggests a possible explanation for shared symptoms. Among these are identity disturbance, feelings of emptiness, self-referential psychotic ideation [48–50], the last of which may be related to excessive, yet often inaccurate, efforts to make sense of other peoples' behavior [6,51–54], i.e. 'over-mentalizing'. Impairments of social functioning in MDD and SCZ patients are often accompanied by reduced hedonic experience [2], which has been associated with blunted reward learning [55,56] and social learning [57,58]. In the present study, social learning parameters were not predicted by ACIPS, a social anhedonia measure.

On the learning level, we observed that deviations in precision weighting (e.g. exagerrated volatility learning in BPD) occurred equally for both information types. As mentioned above, this suggests that aberrant learning occurs independent of domain, in line with previous findings that precision-weighted prediction errors are computed in similar brain regions, irrespective of domain [37,59].

Limitations

We did not use a non-social cue (such as an arrow pointing to a card) as a control condition and therefore cannot fully rule out the possibility that the increased weighting of our social cue observed in BPD and SCZ reflects a more general rather than specifically social peculiarity in information processing. However, eye gaze is a very salient cue and in the paradigm, we aimed to accentuate the social quality of our cue by a clear period of eye contact with the participant before providing the cue.

A further limitation concerns the fact that most patients were in psychopharmacological treatment during data acquisition and had different degrees of disorder severity and chronicity. These variables could not be accounted for with sufficient statistical power in the current sample. Furthermore, different patient groups were assessed in different clinical centers, and there was a gender imbalance in the SCZ and BPD groups.

By adopting a computational psychiatry approach [60–64] to data from an inference task with a social component, we show that BPD patients exhibit an aberrant pattern of learning rate adjustment when the environment becomes more volatile. Instead of quickly relearning changed contingencies, they show exaggerated volatility learning. While SCZ and MDD patients showed a tendency to the same pattern, they did not significantly differ from controls in this respect. We also show that BPD and SCZ patients rely more strongly than controls on social-domain beliefs relative to non-social-domain beliefs when making decisions. Taken together, this shows that there are computational commonalities as well as differences between patient groups, which suggests some underlying mechanisms that may be shared across diagnoses. Since this approach allows for individually quantifying severity of impairment at a mechanistic level, it has the potential to lead to diagnostic and prognostic advances. Furthermore, it points the way to possible targets for novel interventions which transcend traditional diagnostic boundaries.

Supporting information

S1 Text. Supplementary material. The supplement containts contains psychometric and demographic data of our participants, prior configurations of the model parameters, mean posterior estimates of winning model, Full results for ANOVAs using $q(\psi_2)$ and ψ_3.
(DOCX)

Supporting Information Legends

S1 Table. Psychometric data of the participants. All quantities given as Mean ± SD.
S2 Table. Demographic data of the participants. All quantities given as Mean ± SD.
S3 Table. **Prior configurations of perceptual and response model parameters.** Means and variances of Gaussian priors are given in the space in which the parameter was estimated (native, log, or logit).
S4 Table. Mean posterior estimates of learning model and decision model parameters estimated from winning model.
S5 Table. Statistics for mixed ANOVA with averaged $q(\psi_2)$ during stable and volatile phases (Factor Phase) of social and non-social cue (Factor Cue Type) for all groups (Factor Group) and schedules (Factor Schedule).
S6 Table. Statistics for mixed ANOVA with averaged ψ_3 during stable and volatile phases (Factor Phase) of social and non-social cue (Factor Cue Type) for all groups (Factor Group) and schedules (Factor Schedule).

S1 Figure. Learning trajectories for one example participant. A, Precisions ψ_{3card} (red) and ψ_{3gaze} (blue) that modulate the weight on B, prediction errors δ_{2card} (red) and δ_{2gaze} (blue). C, Precision weights ψ_{2card} in red trajectory and $q(\psi_{2card})$ in red dotted trajectory. Precision weights ψ_{2gaze} in blue trajectory and $q(\psi_{2gaze})$ in blue dotted trajectory. Precision weights modulate weight on D) prediction error δ_{1card} (red) and δ_{1gaze} (blue) signals. E, Dark red dots mark the input structure of the non-social information (blue correct=1; green correct=0) and the dotted red line represents the ground truth of this input structure. Light red dots mark the choices (blue card=1; green card=0). The red trajectory is the participant specific belief trajectory about the blue card to be correct that was estimated on the basis of the choices. E, The same logic applies to the social input and response structure in blue. The posterior parameter estimates for this particular participant were ω_{2card} = -1.458, ω_{2gaze} = -3.963, ω_{3card} = -6.056, ω_{2gaze} = -6.05, log(ζ)=-2.623, log(β)=1.477.

S2 Figure. Grouped individual data points showing precision weights for updating beliefs about social and non-social contingency and volatility. A, precision weights $q(\psi_{2)}$. B, precision weights ψ_3. Overall, $q(\psi_2)$ and ψ_3 increase when transitioning from stable to volatile phase.

Acknowledgements

Dr. Andreea Diaconescu was supported by the Swiss National Foundation (PZ00P3_167952) and the Krembil Foundation. Dr. Leonhard Schilbach was supported by the Max Planck Society via a grant for an independent Max Planck Research Group (2015-2020). Lara Henco was supported by the Graduate School of Systemic Neurosciences (LMU). We would like to thank Lea Duerr, Nina von Aken, Mathilda Sommer and Benjamin Pross for helping with the recruitment and data collection.

Author Contributions

Conceptualization: Lara Henco, Leonhard Schilbach, Andreea Diaconescu
Data curation: Lara Henco
Formal analysis: Lara Henco, Christoph Mathys, Andreea Diaconescu, Juha Lahnakoski
Funding acquisition: Leonhard Schilbach
Investigation: Lara Henco, Marie-Luise Brandi, Sophia Hörmann, Johannes Hennings, Irina Papazova, Alkomiet Hasan, Wolfgang Strube
Methodology: Lara Henco, Leonhard Schilbach, Andreea Diaconescu, Christoph Mathys
Project administration: Lara Henco, Leonhard Schilbach
Resources: Leonhard Schilbach, Johannes Hennings, Alkomiet Hasan
Software: Christoph Mathys, Andreea Diaconescu, Juha Lahnakoski
Supervision: Leonhard Schilbach, Christoph Mathys
Validation: Lara Henco, Christoph Mathys, Andreea Diaconescu, Dimitris Bolis
Visualization: Lara Henco
Writing ± original draft: Lara Henco
Writing ± review & editing: Lara Henco, Christoph Mathys, Leonhard Schilbach, Andreea Diaconescu, Juha Lahnakoski, Marie-Luise Brandi, Sophia Hörmann, Johannes Hennings, Irina Papazova, Alkomiet Hasan, Dimitris Bolis, Wolfgang Strube

References

1. Schilbach L. Towards a second-person neuropsychiatry. Philos Trans R Soc B Biol Sci [Internet]. 2015;371(1686):20150011–81. Available from: http://rstb.royalsocietypublishing.org/lookup/doi/10.1098/rstb.2015.0081%0Ahttp://rstb.royalsocietypublishing.org/content/371/1686/20150081
2. Barkus E, Badcock JC. A transdiagnostic perspective on social anhedonia. Front Psychiatry. 2019;10(APR):1–15.
3. Blanchard JJ, Horan WP, Brown SA. Diagnostic differences in social anhedonia: A longitudinal study of schizophrenia and major depressive disorder. J Abnorm Psychol. 2001;110(3):363–71.
4. Kupferberg A, Bicks L, Hasler G. Social functioning in major depressive disorder. Neurosci Biobehav Rev. 2016;69:313–32.
5. Fulford D, Campellone T, Gard DE. Social motivation in schizophrenia : How research on basic reward processes informs and limits our understanding Social motivation in schizophrenia : How research on basic reward processes informs and limits our understanding. Clin Psychol Rev [Internet]. 2018;63(May):12–24. Available from: https://doi.org/10.1016/j.cpr.2018.05.007
6. Frith C. Schizophrenia and theory of mind. Psychol Med. 2004;34:385–9.
7. Fonagy P, Bateman AW. Mentalizing and borderline personality disorder. J Ment Heal [Internet]. 2007 Jan [cited 2012 Oct 27];16(1):83–101. Available from: http://informahealthcare.com/doi/abs/10.1080/09638230601182045
8. Lawson RP, Mathys C, Rees G. Adults with autism overestimate the volatility of the sensory environment. Nat Neurosci [Internet]. 2017;20(9):4–6. Available from: http://www.nature.com/doifinder/10.1038/nn.4615
9. Mathys C. How could we get nosology from computation? Comput Psychiatry New Perspect Ment Illn [Internet]. 2016;20:121–38. Available from: https://mitpress.mit.edu/books/computational-psychiatry
10. Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, et al. Uncertainty in perception and the Hierarchical Gaussian Filter. Front Hum Neurosci [Internet]. 2014;8(November):825. Available from: http://journal.frontiersin.org/article/10.3389/fnhum.2014.00825/abstract
11. Mathys C. A Bayesian foundation for individual learning under uncertainty. Front Hum Neurosci [Internet]. 2011;5(May):1–20. Available from: http://journal.frontiersin.org/article/10.3389/fnhum.2011.00039/abstract

12. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nat Neurosci [Internet]. 2007;10(9):1214–21. Available from: http://www.nature.com/doifinder/10.1038/nn1954

13. Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS. Associative learning of social value. Nature [Internet]. 2008;456(7219):245–9. Available from: http://www.nature.com/doifinder/10.1038/nature07538

14. Sterzer P, Adams RA, Fletcher P, Frith C, Lawrie SM, Muckli L, et al. The Predictive Coding Account of Psychosis. Biol Psychiatry. 2018;1–10.

15. Deserno L, Boehme R, Mathys C, Katthagen T, Kaminski J, Stephan KE, et al. Volatility Estimates Increase Choice Switching and Relate to Prefrontal Activity in Schizophrenia. Biol Psychiatry Cogn Neurosci Neuroimaging [Internet]. 2020 Feb 1;5(2):173–83. Available from: https://doi.org/10.1016/j.bpsc.2019.10.007

16. Petzschner FH, Weber LAE, Gard T, Stephan KE. Computational Psychosomatics and Computational Psychiatry: Toward a Joint Framework for Differential Diagnosis. Biol Psychiatry [Internet]. 2017;82(6):421–30. Available from: http://dx.doi.org/10.1016/j.biopsych.2017.05.012

17. Fineberg SK, Stahl DS, Corlett PR. Computational Psychiatry in Borderline Personality Disorder. Curr Behav Neurosci Reports. 2017;4(1):31–40.

18. Sevgi M, Diaconescu AO, Henco L, Tittgemeyer M, Schilbach L. Social Bayes: Using Bayesian Modeling to Study Autistic Trait–Related Differences in Social Cognition. Biol Psychiatry [Internet]. 2020;87(2):185–93. Available from: https://doi.org/10.1016/j.biopsych.2019.09.032

19. Diaconescu A, Mathys C, Weber LAE, Daunizeau J, Kasper L, Lomakina EI, et al. Inferring on the intentions of others by hierarchical bayesian learning. PLoS Comput Biol [Internet]. 2014;10(9):e1003810. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25187943

20. Diaconescu A, Mathys C, Weber LAE, Kasper L, Mauer J, Stephan KE. Hierarchical prediction errors in midbrain and septum during social learning. Soc Cogn Affect Neurosci. 2017;12(4):618–34.

21. Adams RA, Brown HR, Friston KJ. Bayesian inference , predictive coding and delusions. Avant. 2015;V(3):51–88.

22. Hauser TU, Iannaccone R, Ball J, Mathys C, Brandeis D, Walitza S, et al. Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. JAMA Psychiatry. 2014;71(10):1165–73.

23. Powers AR, Mathys C, Corlett PR. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. Science (80-). 2017;357(August):596–600.

24. DeBerker AO De, Rutledge RB, Mathys C, Marshall L, Cross GF, Dolan RJ, et al. Computations of uncertainty mediate acute stress responses in humans. Nat Commun [Internet]. 2016;7:1–11. Available from: http://dx.doi.org/10.1038/ncomms10996

25. Bernardoni F, Geisler D, King JA, Javadi AH, Ritschel F, Murr J, et al. Altered Medial Frontal Feedback Learning Signals in Anorexia Nervosa. Biol Psychiatry. 2018 Feb 1;83(3):235–43.

26. Sevgi M, Diaconescu AO, Henco L, Tittgemeyer M, Schilbach L. Social Bayes: Using Bayesian modeling to study autistic trait-related differences in social cognition. Biol Psychiatry [Internet]. 2019 Oct 30 [cited 2019 Nov 15]; Available from: https://www.sciencedirect.com/science/article/pii/S0006322319317901?via%3Dihub

27. Bolis D, Schilbach L. Beyond one Bayesian brain: Modeling intra-and inter-personal processes during social interaction: Commentary on "mentalizing homeostasis: The social origins of interoceptive inference" by Fotopoulou & Tsakiris. Neuropsychoanalysis. 2017;19(1):35–8.

28. Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. J Autism Dev Disord. 2001;31(1):5–17.

29. Gooding DC, Pflum MJ. The assessment of interpersonal pleasure: Introduction of the Anticipatory and Consummatory Interpersonal Pleasure Scale (ACIPS) and preliminary findings. Psychiatry Res. 2014;215(1):237–43.

30. Kay SR, Fiszbein A OL. The Positive and Negative Syndrome Scale for schizophrenia. Schizophr Bull. 1987;13(2):261–76.

31. Benkert O, Müller MJ, Schlösser R, Addington D, Wetzel H, Marx-Dannigkeit P. The Calgary Depression Rating Scale for Schizophrenia: development and interrater reliability of a German version (CDSS-G). J Psychiatr Res. 2002;33(5):433–43.

32. Bohus M, Kleindienst N, Limberger MF, Stieglitz RD, Domsalla M, Chapman AL, et al. The short version of the Borderline Symptom List (BSL-23): Development and initial data on psychometric properties. Psychopathology. 2009;42(1):32–9.

33. Stoet G. PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. Teach Psychol. 2017;44(1):24–31.

34. Daunizeau J, den Ouden HEM, Pessiglione M, Kiebel SJ, Stephan KE, Friston KJ. Observing the observer (I): Meta-bayesian models of learning and decision-making.

PLoS One. 2010;5(12).

35. Daunizeau J, den Ouden HEM, Pessiglione M, Kiebel SJ, Friston KJ, Stephan KE. Observing the Observer (II): Deciding When to Decide. PLoS One [Internet]. 2010;5(12):e15555. Available from: http://dx.plos.org/10.1371/journal.pone.0015555

36. Jara-Ettinger J. Theory of mind as inverse reinforcement learning. Curr Opin Behav Sci [Internet]. 2019;29:105–10. Available from: https://doi.org/10.1016/j.cobeha.2019.04.010

37. Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, Ouden HEM Den, et al. Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. 2013;519–30.

38. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage [Internet]. 2009;46(4):1004–17. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1053811909002638

39. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies - revisited. Neuroimage [Internet]. 2014;84:971–85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24018303

40. Sutton RS. Gain Adaptation Beats Least Squares? Proc Seventh Yale Work Adapt Learn Syst [Internet]. 1992;161–166. Available from: papers://d471b97a-e92c-44c2-8562-4efc271c8c1b/Paper/p596

41. Rescorla RA, Wagner AR. A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. 1972;1–18. Available from: papers2://publication/uuid/51EED98C-39D3-4ECA-9CC8-F7E445CCB145

42. Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, denOuden HEM, et al. Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. Neuron [Internet]. 2013;80(2):519–30. Available from: http://dx.doi.org/10.1016/j.neuron.2013.09.009

43. Fineberg SK, Leavitt J, Stahl DS, Kronemer S, Landry CD, Alexander-Bloch A, et al. Differential Valuation and Learning From Social and Nonsocial Cues in Borderline Personality Disorder. Biol Psychiatry [Internet]. 2018;84(11):838–45. Available from: https://doi.org/10.1016/j.biopsych.2018.05.020

44. Gradin VB, Kumar P, Waiter G, Ahearn T, Stickle C, Milders M, et al. Expected value and prediction error abnormalities in depression and schizophrenia. Brain. 2011;134(6):1751–64.

45. Juckel G, Schlagenhauf F, Koslowski M, Wüstenberg T, Villringer A, Knutson B, et al.

Dysfunction of ventral striatal reward prediction in schizophrenia. Neuroimage. 2006;29(2):409–16.

46. Waltz JA, Frank MJ, Robinson BM, Gold JM. Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. 2008;62(7):756–64.

47. Heerey EA, Bell-warren KR, Gold JM. Sensitivity in Schizophrenia. 2009;64(1):62–9.

48. Zandersen M, Parnas J. Exploring schizophrenia spectrum psychopathology in borderline personality disorder. Eur Arch Psychiatry Clin Neurosci [Internet]. 2019;(0123456789). Available from: https://doi.org/10.1007/s00406-019-01039-4

49. Schroeder K, Fisher HL, Schäfer I. Psychotic symptoms in patients with borderline personality disorder. Curr Opin Psychiatry. 2013;26(1):113–9.

50. Debbané M, Salaminios G, Luyten P, Badoud D, Armando M, Tozzi AS, et al. Attachment, neurobiology, and mentalizing along the psychosis continuum. Front Hum Neurosci. 2016;10(August):22.

51. Okruszek Ł, Haman M, Kalinowski K, Talarowska M, Becchio C, Haman M, et al. Impaired Recognition of Communicative Interactions from Biological Motion in Schizophrenia. PLoS One. 2015;10(2):e0116793.

52. Lynch TR, Rosenthal MZ, Kosson DS, Cheavens JS, Lejuez CW, Blair RJR. Heightened sensitivity to facial expressions of emotion in borderline personality disorder. Emotion. 2006;6(4):647–55.

53. Lowyck B, Luyten P, Vanwalleghem D, Vermote R, Mayes LC, Crowley MJ. What's in a face? Mentalizing in borderline personality disorder based on dynamically changing facial expressions. Personal Disord Theory, Res Treat. 2015;7(1):72–9.

54. Corcoran R, Cahill C, Frith CD. The appreciation of visual jokes in people with schizophrenia: A study of "mentalizing" ability. Schizophr Res. 1997;24(3):319–27.

55. Vrieze E, Pizzagalli DA, Demyttenaere K, Hompes T, Sienaert P, De Boer P, et al. Reduced reward learning predicts outcome in major depressive disorder. Biol Psychiatry [Internet]. 2013;73(7):639–45. Available from: http://dx.doi.org/10.1016/j.biopsych.2012.10.014

56. Pizzagalli DA, Iosifescu D, Hallett L, Ratner K, Maurizo F. Reduced Hedonic Capacity in Major Depressive Disorder: Evidence from a Probabilistic Reward Task. J Psychiatr Res. 2009;43(1):76–87.

57. Chevallier C, Tonge N, Safra L, Kahn D, Kohls G, Miller J, et al. Measuring social motivation using signal detection and reward responsiveness. PLoS One. 2016;11(12):1–

14.

58. Safra L, Chevallier C, Palminteri S. Depressive symptoms are associated with blunted reward learning in social contexts. PLoS Comput Biol. 2019;15(7):1–22.

59. Diaconescu AO, Mathys C, Weber LAE, Kasper L, Mauer J, Stephan KE. Hierarchical prediction errors in midbrain and septum during social learning. Soc Cogn Affect Neurosci [Internet]. 2017;(November 2016):nsw171. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28119508%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5390746%5Cnhttps://academic.oup.com/scan/article-lookup/doi/10.1093/scan/nsw171

60. Read Montague P, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. 2013;16(1):72–80.

61. Stephan KE, Mathys C. Computational approaches to psychiatry. Curr Opin Neurobiol [Internet]. 2014;25:85–92. Available from: http://dx.doi.org/10.1016/j.conb.2013.12.007

62. Wang X-J, Krystal JH. Computational Psychiatry. Neuron [Internet]. 2014;84(3):638–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25442941

63. Adams RA, Huys QJM, Roiser JP. Computational Psychiatry: Towards a mathematically informed understanding of mental illness. J Neurol Neurosurg Psychiatry. 2016;87(1):53–63.

64. Huys QJM, Maia T V, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci [Internet]. 2016;19(3):404–13. Available from: http://www.nature.com/reprints/index.html

SUPPLEMENTAL INFORMATION

Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder

Supplemental Tables

S1 Table. Psychometric data of the participants. All quantities given as Mean ± SD.

	HC Participants	**MDD Participants**	**SCZ Participants**	**BPD Participants**	**Significance**
n	**31**	**28**	**29**	**28**	
Gender, (m/f)	(15/16)	(13/15)	(23/6)	(8/20)	$\chi^2(3) = 15.25$, $P = 0.002$ (Chi-Square test)
Age, mean (SD)	35.65 (12.97)	38.43 (10.69)	33.59 (10.01)	31.32 (7.88)	$\chi^2(3) = 5.302$, $P = 0.151$ (Kruskal-Wallis'one way ANOVA)
Years of school, mean (SD)	12.06 (1.99)	12.32 (2.21)	10.87 (3.965)	10.74 (1.973)	F(3,109)= 2.621, $P = 0.054$ (ANOVA)[a]
AQ, mean (SD)	16.66 (5.48)	22.24 (8.23)	21.61 (6.66)	24.72 (6.52)	F(3,103)=7.262, $P <.001$ (ANOVA)[b]
ACIPS, mean (SD)	81.23 (11.75)	68.57 (16.45)	74.11 (15.1)	66 (16.6)	F(3,103)=5.719, $P <.001$ (ANOVA)[b]
CDSS, mean (SD)	-[c]	-	4.11 (4.00)	-	-
PANSS, Positive, mean (SD)	-	-	11.45 (3.43)	-	-
PANSS, Negative, mean (SD)	-	-	13.86 (4.21)	-	-
PANSS, General,	-	-	26.59 (5.82)	-	-

mean (SD)					
PANSS, Total, mean (SD)	-	-	52.24 (11.25)	-	-
BSL-23 (sum), mean (SD)	---[i]	---	---	45.4 (22.59) [d]	

[a] Three missing data points. [b] Nine missing data points. [c] One hyphen indicates that measure applies only to Participants with SCZ. [d] three missing data points.

S2 Table. Demographic data of the participants. All quantities given as Mean ± SD.

	HC Participants	**MDD Participants**	**SCZ Participants**	**BPD Participants**
Age at Diagnosis, mean (SD)	-[a]	28.3(11.62)[b]	26.25 ± 9.82[b]	15.15(5.5)[b]
Number of Hospitalizations, mean (SD)	-	3.14(4.07)	5.88(5.9)[c]	6.14(6)
Duration current Hospitalization (days), mean (SD)	-	20.79(16.63)	65.19(49.74)[d]	13.75(10.71)
Relationship Status No. (%)				
None	12(38.71)	11(39.29)	19(65.52)	14(50)
In a relationship	16(51.61)	4(14.29)	7(24.14)	10(35.71)
Married	2(6.45)	11(39.29)	3(10.35)	2 (7.14)
Divorced	1(3.23)	0(0)	0(0)	2(7.14)
Widowed	0(0)	0(0)	0(0)	0(0)
No answer	0(0)	1(3.57)	0(0)	0(0)
Employment Status No. (%)				
Regularly employed	19(61.3)	13(46.43)	3(10.35)	7(25)
Unemployed	2(6.5)	7(25)	12(41.38)	14(50)
Unable to work	0(0)	3(10.71)	4(13.8)	1(3.6)
Supervised work	0(0)	1(3.571)	3(10.35)	0(0)
Retired	0(0)	1(3.571)	2(6.9)	3(10.71)
In school	10(32.26)	0(0)	3(10.35)	3(10.71)

No answer	0(0)	2(7.24)	2(6.9)	0(0)
Immigration Status **No. (%)**				
Native	28(90.32)	18(64.29)	15(51.72)	18(64.29)
Migrant	3(9.68)	8(28.571)	14(48.28)	9(32.14)
Neuroactive Medications **No. (%)**[d]				
Taking Psychiatric Medications	0(0)	27(96.43)	26(89.66)	26(92.86)
Antidepressants only	0(0)	13(46.43)	0(0)	10(35.71)
Antipsychotics only	0(0)	0(0)	17(58.62)	2(7.14)
Antidepressants and Antipsychotics (Combination)	0(0)	13(46.43)	9(31.03)	13(46.43)
Mood Stabilizer	0(0)	6(21.43)	1(3.45)	1(3.57)
Sedatives	0(0)	0(0)	1(3.45)	2(7.14)
Other	0(0)	3(10.71)	6(20.69)	5(17.86)

[a] One hyphen indicates that the measure only applies to patients. [b] One missing data point. [c] Four missing data points. [d] Thirteen participants with SCZ were not recruited during hospitalization.

S3 Table. **Prior configurations of perceptual and response model parameters.** Means and variances of Gaussian priors are given in the space in which the parameter was estimated (native, log, or logit).

HGF		**Level 1**		**Level 2**		**Level 3**	
Parameter	**Estimation Space**	**Prior Mean**	**Prior Variance**	**Prior Mean**	**Prior Variance**	**Prior Mean**	**Prior Variance**
$\boldsymbol{\mu^{(k=0)}}$	**native**	-	-	0	0	1	0
$\boldsymbol{\sigma^{(k=0)}}$	**log**	-	-	log (0.4)	1	log (0.1)	1
$\boldsymbol{\varphi}$	**logit**	-	-	logit (0)	0	logit (0.1)	2
$\boldsymbol{m}$	**native**	-	-	0	0	1	0
$\boldsymbol{\kappa}$	**log**	log(1)	0	log (1)	0	-	-
$\boldsymbol{\omega}$	**native**	-	-	-4	4	-6	4
ST-K1	**Estimation Space**	Prior Mean	Prior Variance				
$\boldsymbol{\mu}$	**log**	log (1)	0.5				
$\boldsymbol{\hat{r}}$	**log**	log (1)	0				
$\boldsymbol{\hat{v}}$	**logit**	logit (0.5)	0				

h	**logit**	logit (0.005)	1
RW	**Estimation Space**	Prior Mean	Prior Variance
$\nu^{(k=0)}$	**logit**	logit (0.5)	0
α	**logit**	logit (0.5)	1
Response Model	**Estimation Space**	Prior Mean	Prior Variance
ζ	**log**	log (1)	16
β	**log**	log (16)	16

S4 Table. Mean posterior estimates of learning model and decision model parameters estimated from winning model.

Parameter		ω_{2card}	ω_{2gaze}	ω_{3card}	ω_{3gaze}	**log** (ζ)	**log** (β)
mean (SD)	HC	-2.760 (1.805)	-3.142 (1.652)	-5.873 (0.303)	-6.122 (0.376)	-1.324 (2.021)	1.385 (0.650)
	MDD	-2.844 (1.851)	-3.658 (1.513)	-5.986 (0.364)	-6.042 (0.202)	-0.977 (1.644)	0.906 (0.706)
	SCZ	-3.357 (1.147)	-3.802 (2.184)	-5.919 (0.402)	-5.979 (0.177)	0.358 (2.364)	0.620 (1.045)
	BPD	-3.960 (1.491)	-4.891 (2.549)	-5.970 (0.119)	-6.015 (0.142)	0.688 (2.254)	0.932 (1.040)

S5 Table. Statistics for mixed ANOVA with averaged $q(\psi_2)$ during stable and volatile phases (Factor Phase) of social and non-social cue (Factor Cue Type) for all groups (Factor Group) and schedules (Factor Schedule).

Within Subjects Effects			
	df	**F**	**p**
Phase	1	18.628	<.001
Phase x Group	3	0.749	0.526
Phase x Schedule	1	3.988	0.048
Phase x Group x Schedule	3	0.536	0.658
Residual	108		
Information type	1	1.290	0.259

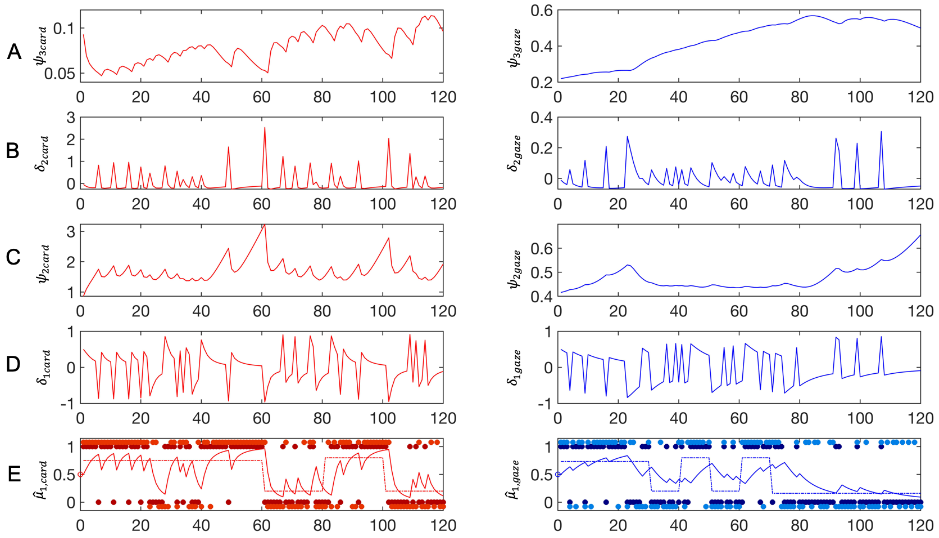
Information type x Group			3	0.946	0.421
Information type x Schedule			1	1.561	0.214
Information type x Group x Schedule			3	0.232	0.874
Residual			108		
Phase x Information type			1	1.654	0.201
Phase x Information type x Group			3	1.644	0.184
Phase x Information type x Schedule			1	1.153	0.285
Phase x Information type x Group x Schedule			3	0.826	0.482
Residual			108		
Between Subjects Effects					
Group			3	3.939	0.01
Schedule			1	0.936	0.335
Group x Schedule			3	0.661	0.578
Residual			108		
Post Hoc Tests					
			t	**Cohens's d**	**pbonf**
Phase	**stable**	**volatile**	-4.419	-0.410	<.001
Group	**HC**	**MDD**	1.023	0.095	1.000
		SCZ	1.803	0.167	0.445
		BPD	3.346	0.311	0.007
	MDD	**SCZ**	0.749	0.070	1.000
		BPD	2.264	0.210	0.154
	SCZ	**BPD**	1.537	0.143	0.764

S6 Table. Statistics for mixed ANOVA with averaged ψ_3 during stable and volatile phases (Factor Phase) of social and non-social cue (Factor Cue Type) for all groups (Factor Group) and schedules (Factor Schedule).

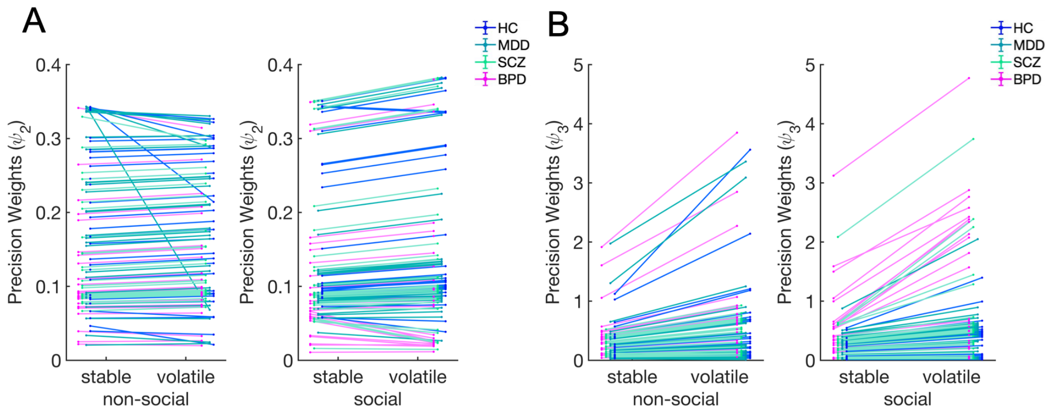
Within Subjects Effects			
	df	**F**	**p**
Phase	1	125.990	< .001
Phase x Group	3	6.980	< .001

Phase x Schedule			1	3.340	0.070
Phase x Group x Schedule			3	2.207	0.091
Residual			108		
Information type			1	0.810	0.370
Information type x Group			3	1.220	0.306
Information type x Schedule			1	1.046	0.309
Information type x Group x Schedule			3	0.380	0.768
Residual			108		
Phase x Information type			1	2.188	0.142
Phase x Information type x Group			3	2.625	0.054
Phase x Information type x Schedule			1	1.907	0.170
Phase x Information type x Group x Schedule			3	0.336	0.800
Residual			108		
Between Subjects Effects					
Group			3	7.159	< .001
Schedule			1	5.118	0.026
Group x Schedule			3	2.530	0.061
Residual			108		
Post Hoc Tests					
			t	**Cohens's d**	**pbonf**
Phase	**stable**	**volatile**	-10.06	-0.934	<.001
Group	**HC**	**MDD**	-0.731	-0.068	1.000
		SCZ	-1.081	-0.100	1.000
		BPD	-4.332	-0.402	<.001
	MDD	**SCZ**	-0.334	-0.031	1.000
		BPD	-3.510	-0.326	0.004
	SCZ	**BPD**	-3.210	-0.298	0.010

Supplemental figures



S1 Figure. Learning trajectories for one example participant. A, Precisions ψ_{3card} (red) and ψ_{3gaze} (blue) that modulate the weight on B, prediction errors δ_{2card} (red) and δ_{2gaze} (blue). C, Precision weights ψ_{2card} in red trajectory and $q(\psi_{2card})$ in red dotted trajectory. Precision weights ψ_{2gaze} in blue trajectory and $q(\psi_{2gaze})$ in blue dotted trajectory. Precision weights modulate weight on D) prediction error δ_{1card} (red) and δ_{1gaze} (blue) signals. E, Dark red dots mark the input structure of the non-social information (blue correct=1; green correct=0) and the dotted red line represents the ground truth of this input structure. Light red dots mark the choices (blue card=1; green card=0). The red trajectory is the participant specific belief trajectory about the blue card to be correct that was estimated on the basis of the choices. E, The same logic applies to the social input and response structure in blue. The posterior parameter estimates for this particular participant were ω_{2card} = -1.458, ω_{2gaze} = -3.963, ω_{3card} = -6.056, ω_{2gaze} = -6.05, log(ζ)=-2.623, log(β)=1.477.



S2 Figure. Grouped individual data points showing precision weights for updating beliefs about social and non-social contingency and volatility. A, precision weights $q(\psi_2)$. B, precision weights ψ_3. Overall, $q(\psi_2)$ and ψ_3 increase when transitioning from stable to volatile phase.

4 Discussion

This thesis adopted a Bayesian modelling approach to behaviour obtained from a previously established probabilistic reward learning task (Sevgi et al., 2020) that involved learning of social and non-social information. Using computational modelling of behaviour, we estimated individual *learning and decision-making fingerprints* that reflect individual cognitive variation when learning about social and non-social aspects of the environment.

The first part of the discussion will present the results of the computational model comparison, model validity analyses and simulations to elucidate the computational mechanisms of the integration of social information in decision-making and the individual propensity to make use of social information during the learning task.

I will then present findings from the first study (chapter 2) that used fMRI to uncover the neural activity associated with social and non-social predictions and the inter-individual variation in the propensity to weight social over non-social predictions. Following this, results of the second study (chapter 3) will be presented, in which these processes were investigated from behaviour in patients with BPD, SCZ and MDD. Relating these studies to each other, potential neural signatures of aberrant social weighting in patients with BPD and SCZ will be discussed in light of the previous literature.

The third part looks at the neural activations associated with social and non-social outcome processing, i.e. prediction errors. Results of the behavioural psychiatric patient study will then be discussed in which I identified aberrant learning for both social and non-social information in BPD and will relate these to the fMRI results in order to generate hypotheses for potential underlying neural mechanisms.

Finally, I discuss methodological and interpretational considerations for future studies.

4.1 Bayesian modelling of learning and decision-making

4.1.1 Model comparison in fMRI and behavioural patient study

The work presented in this thesis fitted different reinforcement and Bayesian learning models to participants behaviour. Model comparison in both studies revealed that the three-level HGF best explained participants behaviour in the task, which is in line with previous studies on learning under uncertainty that indicate a clear superiority of the HGF over other, traditional reinforcement learning models such as the Rescorla Wagner learning model (Bernardoni et al., 2018; DeBerker et al., 2016; Diaconescu et al., 2017; Iglesias et al., 2013). This suggests that participants inferred upon the volatility of the card and the gaze information in order to predict the outcome of the task. In the patient study, I additionally ran posterior predictive analyses to further account for the robustness of the model. These analyses showed that responses simulated from posterior estimates produced the same group differences in performance, i.e. significantly lower performance in SCZ and BPD patients compared to controls.

4.1.2 Validation analyses for weighting advice accuracy during decision-making

A central question in both studies concerned the integration of social information during decision-making. To this end, I estimated the individual tendency to make use of social information, both computationally and by means of model-agnostic measures. In the fMRI study, we conducted proof-of-concept analyses that showed that the computational parameter ζ was positively correlated with answers of post-experimental questionnaire indicating how much participants used the gaze during the task. In addition, eye-tracking during scanning showed that the parameter was significantly correlated with fixations falling on the computer-generated face during the decision-making period. Additionally, simulation analyses conducted in both studies showed that agents adopting high ζ show a greater sensitivity to the social input over the non-social input, while low ζ values indicate greater sensitivity with regard to the non-social input. Thus, the model was capable of capturing an individual *social learning and decision-making fingerprint.*

4.2 Tracking and weighting advice accuracy

4.2.1 Neural correlates of tracking and weighting advice accuracy

The fMRI study aimed to investigate the neural correlates of the predicted advice accuracy, i.e. probability of the gaze to give a correct advice ($\hat{\mu}_{1,gaze}^{(t)}$), during decision-making and the inter-individual variability of these. Tracking social accuracy $\hat{\mu}_{1,gaze}^{(t)}$ was associated with activity in the inferior temporal gyri, inferior and superior parietal lobule as well as parts of the striatum including the right putamen and pallidum. In addition, the results demonstrated that participants who indicated using the social cue more, showed greater activity in response to $\hat{\mu}_{1,gaze}^{(t)}$ in bilateral putamen and anterior insula.

The involvement of the putamen in tracking the social-domain belief conjugates with previous studies investigating the neural correlates involved in learning about others (Báez-Mendoza & Schultz, 2013). For instance, Delgado, Frank, & Phelps (2005) showed that the putamen, insula as well as the vSTR showed stronger activity when participants decided to share rather than keep their money during iterated trust games, i.e. when reciprocation is predicted. Similarly, King-Casas et al. (2005) found that the caudate activity was associated with trust decisions in a trust game.

Yet, the striatum is not uniquely involved in the computations of social value but also plays a crucial role in other forms of (non-social) reward based-learning (O'Doherty, 2004). The generic role of the striatum in value learning has further been supported by a recent meta-analysis (Gu et al., 2019) that compared the neural substrates of social and monetary reward anticipation. This analysis revealed that the striatum together with the insula, but also VTA and SMA play an important role in the valuation and anticipation of social *and* non-social reward. Interestingly however, a differential contrast looking at social compared to monetary reward anticipation revealed more consistent activation in the putamen and dorsal anterior insula, whereas the reverse contrast showed more consistent activation for the VS, dorsal anterior cingulate cortex (dACC) and ventral anterior insula for monetary rewards. In addition, increased co-activation of the insula and putamen in response to social compared to non-social stimuli has previously been shown in spatial-cueing tasks comparing gaze and arrow cues (Greene et al., 2011).

Hence, although the putamen and insula are involved in non-social processing, these findings suggest that social stimuli can engage these to a greater extent. The finding that the putamen and insula were correlated with increased weighting of social-domain predictions during decision-making is especially intriguing given the results of our transdiagnostic patients study.

4.2.2 Weighting advice accuracy in BPD and SCZ

The second behavioural study demonstrated that patients with SCZ and BPD weighted the predicted advice accuracy ($\hat{\mu}_{1,gaze}^{(t)}$) significantly more compared to HC and patients with MDD. Increased weighting of social information is in line with findings showing that patients with SCZ and BPD both show a hyper-sensitivity to social cues, such as faces and gaze (Berchio et al., 2017; Langdon et al., 2017). For instance, studies investigating the automatic processing of and orienting to social gaze have shown that patients with SCZ show stronger gaze cueing effects and an impaired disengagement from a location cued by gaze (Langdon et al., 2017). In addition, social hyper-sensitivity in SCZ has been associated with the exaggerated need to make sense of other people's minds and inferring intentionality (Frith, 2004; cf. next section). Alternatively, a large body of evidence suggests a general mechanism of salience over-attribution to neutral cues (Maia & Frank, 2017), which could be an alternative explanation for increased cue using. Although the hyper-mentalization and aberrant salience accounts are not mutually exclusive, future studies should employ designs using social and non-social cues to fully delineate whether the present findings can be attributed to *social* decision-making alone. The hyper-sensitivity account of BPD is mostly related to heightened expectation of negative social events and general mistrust in social partners (Gunderson et al., 2018; Herpertz & Bertsch, 2015). Since our study did not involve emotional but neutral gaze cues, enhanced social weighting in BPD may suggest that the mere presence of a neutral social cue can induce the sensitivity bias that has previously been observed for emotional stimuli. In fact, fear of abandonment and rejection in BPD may induce a general excessive need to predict others thoughts and actions, i.e. engage in excessive mentalization. Thus, the common mechanism of increased social cue weighting in BPD and SCZ may be subserved by a joint disposition of increased mental state attribution.

4.2.3 Potential neural fingerprints of weighting advice accuracy in BPD and SCZ

The fMRI study (chapter 2) demonstrated that HC who tracked the social cue more, showed increased activity in the insula and putamen. Considering our behavioural results of enhanced social weighting in BPD and SCZ, I would consequently predict that this would be reflected in enhanced insula and putamen activity.

Moreover, considering previous evidence on hyper-mentalization in BPD and SCZ, I would also predict that the neural correlates of tracking the social cue in these disorders would additionally involve mentalization areas such as TPJ, STS and dmPFC, which were not significantly active in our study with HC. Thus, a significant group difference between patients with BPD and SCZ and HC in 'mentalization areas' can be predicted due to a hyper-intentionality attributed to the computer-generated face. Previously, studies have demonstrated that BOLD activity in reward-processing and mentalization areas were significantly modulated by the amount of intentionality attributed to a confederate (Rilling et al., 2002; Singer, Kiebel, Winston, Dolan, & Frith, 2004). For instance, activity in the STS was significantly stronger in participants that were told to be playing with intentional compared to unintentional agents in the prisoners dilemma game (Singer et al., 2004). Thus, it is possible that patients with BPD and SCZ automatically infer intentionality activating those brain regions involved in inferential processing, without explicit instructions to do so.

Mounting evidence supports this presumption as SCZ patients are known to attribute intentionality to neutral cues (Frith, 2004). For example, (Okruszek et al., 2015) showed that participants with SCZ tend to misattribute non-communicative actions and gestures as communicative. In a similar vein, Walter et al. (2009) conducted a study in which healthy and

SCZ participants were asked to choose the appropriate ending to stories containing intentional or physical causation. During intentional inference, HC as well as SCZ patients showed increased activity in the medial PFC and STS, which have previously been involved in mentalization. During physical causation, activity in these regions decreased significantly in HC but not in SCZ patients. The finding that regions associated with mentalization were equally engaged in the non-intentional inference in patients with SCZ has been replicated (Ciaramidaro et al., 2015) and further converges with the notion of over-mentalizing in SCZ (Corcoran, Cahill, & Frith, 1997; Frith, 2004; Okruszek et al., 2015).

Similarly, patients with BPD have been found to show increased activity in mentalization areas in fMRI studies. Recently, it was demonstrated that compared to HC, patients with BPD showed greater activity to social anticipatory cues in the form of emotional faces, which was associated with more pronounced activity in the STS (Doell et al., 2019). This is in line with the elevated need to predict other people's behaviour.

4.3 Social and non-social belief-updating

4.3.1 Neural correlates of social and non-social prediction errors

The fMRI analysis showed that negative social prediction errors ($\delta_{1gaze}^{(t)} < 0$, i.e., gaze misleading) activated the right insula, rolandic operculum and left posterior medial frontal gyrus during wrong outcomes but not during correct choices. This suggests that in our task, wrong advice was especially salient, when the outcome was wrong (i.e. the participant followed the gaze but shouldn't) and can be understood as a warning signal. The bilateral insula was equally implicated in non-social prediction errors $\delta_{1card}^{(t)}$ (i.e. surprise about winning card colour), supporting its generic role in monitoring error and risk (Bossaerts, 2010; d'Acremont, Lu, Li, Van der Linden, & Bechara, 2009; Diaconescu et al., 2017). The insula is functionally coupled with the anterior cingulate cortex and constitute the primary components of the salience network that is involved in the evaluation and selection of highly salient stimuli relevant for goal directed behaviour.

4.3.2 Social and non-social learning in BPD

The behavioural patient study showed that patients with BPD differed significantly from HC in how they learned from social and non-social outcomes. Given the pivotal role of social impairments in BPD, it is surprising that non-social learning was equally affected in this disorder. However, our finding is in line with a previous study of Fineberg et al. (2018) that showed reduced learning from social and non-social cues in a probabilistic task. In predictive coding terms, this learning impairment would refer to relatively precise priors at lower levels of the processing hierarchy and that the errors pertaining those priors are not used to update the predictions accordingly. In contrast to blunted probability learning, we found that patients with BPD showed excessive learning for social and non-social volatility, which means decreased high-level precision of prior beliefs about the stability of the environment.

4.3.3 Potential neural fingerprints of aberrant learning in BPD

In neural terms, I would predict that impaired probability learning in BPD would be associated with blunted activity in the insula, that signalled both social and non-social prediction errors in our fMRI task.

While some evidence indicates increased activation or impaired deactivation in the insula and amygdala, associated with a negativity bias in emotion processing (Ducasse, Courtet, & Olié, 2014; Herpertz & Bertsch, 2015), social interaction tasks have painted a more nuanced picture: In the seminal work of King-Casas et al. (2008), HC and patients with BPD acted as trustees in a trust game. When HC received small offers, they showed a tendency to repay generous amounts back to the investor as an attempt to repair and maintain the cooperation. By contrast, patients with BPD repaid smaller amounts causing a rupture in the cooperation. Moreover, using fMRI, the study demonstrated that HC showed significantly greater activity in the anterior insula in response to low vs. high investments, which was not found in patients with BPD. However, during the repayment phase, both HC and BPD showed increased insula activity. Thus, the absent insular activation in response to small investments was interpreted as an impaired ability to register social norm violations due to default prior expectations of negative social experience. Accordingly, negative social outcomes are less surprising than positive outcomes to patients with BPD. This is in line with another study adopting a virtual ball tossing game in which participants are either socially excluded, included or neither (control) (Domsalla et al., 2014). Interestingly, while patients felt just as excluded as HC in the exclusion condition, they felt significantly more excluded in the inclusion and control condition. In addition, while HC showed insula activity modulated by exclusion, this was not shown in patients with BPD, which converges with the findings of King-Casas et al. (2008).

Thus, according to these results, we may predict blunted insular activation for $\delta_{1gaze}^{(t)}$ in BPD compared to controls. However, the previous results explained the absence of insula activity in social norm violation by a match between outcome and prediction. Thus, modelling the prediction error as I did in the current project, may yield different findings since the surprise signal is computed under consideration of the prediction.

With regard to possible neural correlates of excessive volatility learning in BPD, previous studies demonstrated an involvement in volatility learning in the ACC, the septum, which is part of the cholinergic midbrain, and the dlPFC (Behrens et al., 2008; Deserno et al., 2020; Diaconescu et al., 2017; Iglesias et al., 2013). Interestingly, Domsalla et al. (2014) found that across all conditions in the ball tossing game, patients with BPD showed an unspecific hyperactivity of the dACC and dlPFC and dmPFC, suggesting that these error-monitoring areas fail to be regulated in positive situations. Speculatively, one could suggest that the rather diffuse hyperactivity of the dACC reflects heightened volatility in the environment which causes an inability to learn from social situations and differentiate positive from negative situations. This would be reminiscent of our findings of increased volatility learning at the expense of compromised contingency learning.

4.3.4 Social and non-social learning in SCZ and MDD

In our behavioural study, we did not find significant impairments in learning about the probabilistic contingencies of social or non-social cues in patients with SCZ. This is in contrast to previous studies showing impaired probability learning (Culbreth et al., 2016; Schlagenhauf et al., 2014) and suggests that in our tasks, patients with SCZ did not show significantly altered priors as previously suggested (Adams, Stephan, Brown, Frith, & Friston, 2013; Diaconescu et al., 2019; Schmack, Rothkirch, Priller, & Sterzer, 2017; Sterzer, Voss, et al., 2018). For instance, Powers, Mathys, & Corlett (2017) demonstrated that decision-making in individuals with auditory hallucinations was characterized by increased weighting of the prior belief over sensory data. However, there is also evidence suggesting reduced prior precision in patients with SCZ (Jardri, Duverne, Litvinova, & Dene, 2017; Schmack et al., 2017; Stuke, Weilnhammer, Sterzer, & Schmack, 2019). To accommodate these ostensibly equivocal findings, previous suggestions have pointed to an imbalance of prior precision across different levels of the processing hierarchy (Diaconescu et al., 2019; Sterzer, Adams, et al., 2018). Adopting a hierarchical model in the behavioural study, we did find a tendency of increased high-level volatility learning in patients with SCZ, however it did not significantly differ from controls. Previously, Deserno et al. (2020) found that patients with SCZ demonstrated aberrant initial beliefs pertaining the volatility of the environment, rather than an impaired evolution of this belief in time Deserno et al. (2020). Future studies shall disentangle the role of aberrant high-order initial beliefs and its propagation in SCZ and also investigate the impact of model choice and prior settings on these seemingly divergent findings.

One reason for not finding aberrant learning in SCZ in chapter 3 may be that precision weights were analysed during stable and volatile phases. Instead, it is possible that patients with SCZ rather show an imbalance in precision weights between positive and negative (social) prediction errors. For instance, in social interactions it is possible that negative advice is weighted more strongly due to negatively biased predictions about the intention of others (Diaconescu et al., 2019).

Similarly, there is well-established evidence that demonstrates a negativity bias in MDD, which is characterised by preferential processing and heightened sensitivity to negative (social) experiences (Kupferberg et al., 2016). Thus, it is possible that patients with MDD grant more precision to negative outcomes than positive outcomes during learning (Pulcu & Browning, 2019). In fact, a previous computational study showed that precision of positive and negative outcomes are tracked independently and could, in the case of MDD, be biasedly processed (Pulcu & Browning, 2017). In line with this notion, a recent study, which compared performance of patients with MDD in nine different decision-making tasks (Mukherjee, Lee, Kazinka, D Satterthwaite, & Kable, 2020), demonstrated that the tasks which require learning to avoid punishment, revealed the most profound impairments in decision-making in patients with MDD.

The paradigm used in this thesis, in contrast to previous studies showing aberrant reward learning (cf. Rothkirch, Tonn, Köhler, & Sterzer, 2017), did not include punishment for wrong choices (points were scored for correct choices but not deducted for incorrect choices). This may at least partially explain our negative findings with regard to MDD and SCZ. One further explanation concerns the heterogeneity of our MDD sample. In our study, we did not distinguish between subtypes, such as melancholic or atypical MDD, the latter of which exhibits intact mood reactivity to positive events. In fact, some evidence suggests that aberrant reward learning and its associated neural fingerprint of reduced activity in the vSTR is only observed in patients with impaired mood reactivity (Foti, Carlson, Sauder, & Proudfit, 2014).

4.4 Future directions

4.4.1 Methodological considerations

Using the present approach of uninstructed social inference, we were able to probe inter-individual differences in automatic and spontaneous social inference. To this end, we adopted a computer-generated face giving implicit advice by means of gaze cues. Gaze cues are important sources for mental state attribution and have been found to be differentially processed compared to non-social cues (e.g. Greene et al., 2011). Especially brain structures involved in mentalization, such as the TPJ and STS respond to direct gaze cues (Senju & Johnson, 2009). Similarly, we found that BOLD activity in the superior temporal gyrus significantly correlated with the number of fixations on the face during decision-making. This supports the finding that the mere presence of direct gaze triggers mentalization processes.

Although the extent to which participants weighted the social cue, as indicated by the computational parameter, was significantly associated with face fixations during decision-making, the computational parameters did not significantly correlate with activity in mentalization areas, which is in contrast to previous studies probing instructed mental state attribution (Behrens et al., 2008; Diaconescu et al., 2017). Previous studies have demonstrated that BOLD activity in mentalization areas were significantly stronger when participants were explicitly told to engage with an intentional agent (Rilling et al., 2002; Singer et al., 2004). Therefore, the absence of information about the computer-generated face could have affected the degree to which participants engaged in mentalization.

A further methodological consideration regards the lack of a non-social cue as a control condition, which is why we cannot fully delineate whether the present findings pertain to more general as opposed to specifically 'social' cueing effects. However, there are various methodological difficulties associated with using a non-social control cue. Within-subject designs would give rise to a number of challenges: A repeated exposure to the same probabilistic schedule would yield improved performance and different parameter estimations due to rather unspecific learning effects. On the other hand, employing two different schedules for two conditions would also have an impact on parameter estimation. This could only be solved in a counter-balanced presentation of both schedules, requiring a much larger pool of participants. Alternatively, one could implement a between-subjects design, that would however introduce uncontrolled individual differences. Another issue regarding the present and previous paradigms (Behrens et al., 2008; Diaconescu et al., 2017) is that the valence of the social advice has always been defined with respect to the monetary outcome, whereby helpful advice leads to monetary rewards and misleading advice to monetary losses. Thus, the social processes might be considered confounded by other reward related processes (Bellucci, Molter, & Park, 2019). One possibility would be to include social feedback in the form of positive face expressions for correct choices or and negative face expressions for negative choices.

4.4.2 Outlook

While we interpreted the commonality of over-weighting social-domain predictions in BPD and SCZ patients in light of a common disposition to hyper-mentalize, it should be noted that the current data cannot rule out alternative explanations. Future studies should address this by combining our paradigm with other, established paradigms of mentalization such as the Strange Stories and Strange Cartoons Task. Moreover, conducting fMRI studies in patients investigating the hypotheses outlined in 2.5, will further investigate whether this commonality is also subserved by the activation of mentalization areas.

In addition, hyper-mentalization in SCZ has been associated with self-referential thinking and psychotic ideation (Frith, 2004). Since a significant proportion of patients with BPD experience psychotic symptoms (Debbané et al., 2016; Schroeder, Fisher, & Schäfer, 2013; Zandersen & Parnas, 2019), it seems plausible that hyper-mentalization in our task as indicated by over-weighting of social information could be attributed to psychotic ideation.

Moreover, to further establish the relationship between the computational estimates and the ability to mentalize, we are currently assessing behaviour of patients with autistic spectrum disorder (ASD) obtained from our inference task. In contrast to SCZ and BPD patients, ASD patients show a reduced tendency to engage in mentalization, especially in tasks that probe spontaneous and automatic mental inference (Senju et al., 2009). The differences in the computational parameters between these patient groups could help shed light on the mechanistic foundations of these opposing social symptoms as well as the disorders themselves. Specifically, I hypothesise that hypo-mentalization will be reflected in a reduced tendency to weight social-domain predictions and we will assess the role of the oxytocin release system on social inference in patients with ASD and HC. Oxytocin is a neuropeptide that has been shown to modulate mentalizing abilities (Domes, Heinrichs, Michel, Berger, & Herpertz, 2007) and may therefore be an interesting candidate to investigate the neuroendocrinological mechanisms involved in the computations of underlying (aberrant) spontaneous mental state inference. Thus, the modelling approach presented in this thesis in combination with physiological and neuroendocrine measures may help to develop computationally informed biomarkers that may be useful for understanding the impairments in social interaction in psychiatric patients and may offer objective and mechanistic progress indicators for therapeutic interventions.

5 Conclusions

The present thesis used computational modelling to identify the underlying mechanisms involved in social and non-social learning and decision-making. More specifically, this thesis aimed to improve our understanding of the neural correlates of uninstructed social inference and inter-individual differences in the spontaneous integration of social information during decision-making. In addition, the thesis adopted a transdiagnostic computational psychiatry approach to investigate these processes in psychiatric patients who show fundamental impairments in social cognition. The findings demonstrated computational commonalities that cut across diagnostic boundaries and may relate to transdiagnostic impairments in mentalization, as well as computational features that distinguished between patient groups. Thus, the projects computational phenotyping approach allows for an objective estimation of individual *social learning and decision-making signatures,* which could be highly relevant for diagnostic, prognostic and therapeutic advances in clinical psychiatry.

6 References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The Computational Anatomy of Psychosis. *Frontiers in Psychiatry*, *4*(May), 1–26. https://doi.org/10.3389/fpsyt.2013.00047

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington DC.

Báez-Mendoza, R., & Schultz, W. (2013). The role of the striatum in social behavior. *Frontiers in Neuroscience*, *7*(7 DEC), 1–14. https://doi.org/10.3389/fnins.2013.00233

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “Reading the Mind in the Eyes” Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251. https://doi.org/10.1017/S0021963001006643

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. https://doi.org/10.1038/nature07538

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Behrens, T., Hunt, L., & Rushworth, M. (2009). The Computation of Social Behavior. *Science*, *324*(5931), 1160–1164. https://doi.org/10.1126/science.1169694

Bellucci, G., Molter, F., & Park, S. Q. (2019). Neural representations of honesty predict future trust behavior. *Nature Communications*, *10*(1), 1–12. https://doi.org/10.1038/s41467-019-13261-8

Berchio, C., Piguet, C., Gentsch, K., Küng, A. L., Rihs, T. A., Hasler, R., … Perroud, N. (2017). Face and gaze perception in borderline personality disorder: An electrical neuroimaging study. *Psychiatry Research - Neuroimaging*, *269*(March), 62–72. https://doi.org/10.1016/j.pscychresns.2017.08.011

Berlin, H. A., Rolls, E. T., & Iversen, S. D. (2005). Borderline Personality Disorder, Impulsivity, and the Orbitofrontal Cortex. *American Journal of Psychiatry*, *162*(12), 2360–2373. https://doi.org/10.1176/appi.ajp.162.12.2360

Bernardoni, F., Geisler, D., King, J. A., Javadi, A. H., Ritschel, F., Murr, J., … Ehrlich, S. (2018). Altered Medial Frontal Feedback Learning Signals in Anorexia Nervosa. *Biological Psychiatry*, *83*(3), 235–243. https://doi.org/10.1016/j.biopsych.2017.07.024

Bossaerts, P. (2010). Risk and risk prediction error signals in anterior insula. *Brain Structure & Function*, *214*(5–6), 645–653. https://doi.org/10.1007/s00429-010-0253-1

Chekroud, A. M. (2015). Unifying treatments for depression: An application of the Free Energy Principle. *Frontiers in Psychology*, *6*(FEB), 1–8. https://doi.org/10.3389/fpsyg.2015.00153

Ciaramidaro, A., Bölte, S., Schlitt, S., Hainz, D., Poustka, F., Weber, B., … Walter, H. (2015). Schizophrenia and autism AS contrasting minds: Neural evidence for the hypo-hyper-intentionality hypothesis. *Schizophrenia Bulletin*, *41*(1), 171–179. https://doi.org/10.1093/schbul/sbu124

Clark, J. E., Watson, S., & Friston, K. J. (2018). What is mood? A computational perspective. *Psychological Medicine*, *48*(14), 2277–2284. https://doi.org/10.1017/S0033291718000430

Corcoran, R., Cahill, C., & Frith, C. D. (1997). The appreciation of visual jokes in people with schizophrenia: A study of “mentalizing” ability. *Schizophrenia Research*, *24*(3), 319–327. https://doi.org/10.1016/S0920-9964(96)00117-X

Culbreth, A. J., Gold, J. M., Cools, R., & Barch, D. M. (2016). Impaired activation in cognitive control regions predicts reversal learning in schizophrenia. *Schizophrenia Bulletin*, *42*(2), 484–493. https://doi.org/10.1093/schbul/sbv075

d’Acremont, M., Lu, Z. L., Li, X., Van der Linden, M., & Bechara, A. (2009). Neural correlates of risk prediction error during reinforcement learning in humans. *NeuroImage*, *47*(4), 1929–1939. https://doi.org/10.1016/j.neuroimage.2009.04.096

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*(2), 199–204. https://doi.org/10.1016/j.conb.2006.03.006

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453. https://doi.org/10.3758/CABN.8.4.429

Debbané, M., Salaminios, G., Luyten, P., Badoud, D., Armando, M., Tozzi, A. S., … Brent, B. K. (2016). Attachment, neurobiology, and mentalizing along the psychosis continuum. *Frontiers in Human Neuroscience*, *10*(August), 22. https://doi.org/10.3389/fnhum.2016.00406

DeBerker, A. O. De, Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*, *7*, 1–11. https://doi.org/10.1038/ncomms10996

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611–1618. https://doi.org/10.1038/nn1575

Delgado, Mauricio R., Li, J., Schiller, D., & Phelps, E. A. (2008). Review. The role of the striatum in aversive learning and aversive prediction errors. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3787–3800. https://doi.org/10.1098/rstb.2008.0161

Delgado, Mauricio R., Nearing, K. I., LeDoux, J. E., & Phelps, E. A. (2008). Neural Circuitry Underlying the Regulation of Conditioned Fear and Its Relation to Extinction. *Neuron*, *59*(5), 829–838. https://doi.org/10.1016/j.neuron.2008.06.029

Deserno, L., Boehme, R., Mathys, C., Katthagen, T., Kaminski, J., Stephan, K. E., … Schlagenhauf, F. (2020). Volatility Estimates Increase Choice Switching and Relate to Prefrontal Activity in Schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(2), 173–183. https://doi.org/10.1016/j.bpsc.2019.10.007

Diaconescu, A., Hauke, D. J., & Borgwardt, S. (2019). Models of persecutory delusions: a mechanistic insight into the early stages of psychosis. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-019-0427-z

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

Diaconescu, A., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, *12*(4), 618–634. https://doi.org/10.1093/scan/nsw171

Dixon-Gordon, K. L., Tull, M. T., Hackel, L. M., & Gratz, K. L. (2017). The Influence of Emotional State on Learning From Reward and Punishment in Borderline Personality Disorder. *Journal of Personality Disorders*, *32*(4), 433–446. https://doi.org/10.1521/pedi_2017_31_299

Doell, K. C., Olié, E., Corradi-Dell'Acqua, C., Courtet, P., Perroud, N., & Schwartz, S. (2019). Atypical processing of social anticipation and feedback in borderline personality disorder. *NeuroImage: Clinical*, *25*(May 2019), 102126. https://doi.org/10.1016/j.nicl.2019.102126

Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007). Oxytocin

Improves “Mind-Reading” in Humans. *Biological Psychiatry*, *61*(6), 731–733. https://doi.org/10.1016/j.biopsych.2006.07.015

Domes, G., Schulze, L., & Herpertz, S. C. (2009). Emotion recognition in borderline personality disorder-a review of the literature. *Journal of Personality Disorders*, *23*(1), 6–19. https://doi.org/10.1521/pedi.2009.23.1.6

Domsalla, M., Koppe, G., Niedtfeld, I., Vollstädt-Klein, S., Schmahl, C., Bohus, M., & Lis, S. (2014). Cerebral processing of social rejection in patients with borderline personality disorder. *Social Cognitive and Affective Neuroscience*, *9*(11), 1789–1797. https://doi.org/10.1093/scan/nst176

Ducasse, D., Courtet, P., & Olié, E. (2014). Physical and social pains in borderline disorder and neuroanatomical correlates: A systematic review. *Current Psychiatry Reports*, *16*(5). https://doi.org/10.1007/s11920-014-0443-2

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*(5643), 290–292. https://doi.org/10.1126/science.1089134

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience*, *6*(OCT), 1–17. https://doi.org/10.3389/fnins.2012.00148

Fertuck, E. a, Jekal, a, Song, I., Wyman, B., Morris, M. C., Wilson, S. T., … Stanley, B. (2009). Enhanced’Reading the Mind in the Eyes’ in borderline personality disorder compared to healthy controls. *Psychological Medicine*, *39*(12), 1979. https://doi.org/10.1017/S003329170900600X.Enhanced

Fineberg, S. K., Leavitt, J., Stahl, D. S., Kronemer, S., Landry, C. D., Alexander-Bloch, A., … Corlett, P. R. (2018). Differential Valuation and Learning From Social and Nonsocial Cues in Borderline Personality Disorder. *Biological Psychiatry*, *84*(11), 838–845. https://doi.org/10.1016/j.biopsych.2018.05.020

Fonagy, P., Luyten, P., & Bateman, A. (2015). *Translation : Mentalizing as Treatment Target in Borderline Personality Disorder*. *6*(4), 380–392.

Foti, D., Carlson, J. M., Sauder, C. L., & Proudfit, G. H. (2014). Reward dysfunction in major depression: Multimodal neuroimaging evidence for refining the melancholic phenotype. *NeuroImage*, *101*, 50–58. https://doi.org/10.1016/j.neuroimage.2014.06.058

Frick, C., Lang, S., Kotchoubey, B., Sieswerda, S., Dinu-Biringer, R., Berger, M., … Barnow, S. (2012). Hypersensitivity in borderline personality disorder during mindreading. *PLoS ONE*, *7*(8), 4–11. https://doi.org/10.1371/journal.pone.0041650

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148–158. https://doi.org/10.1016/S2215-0366(14)70275-5

Frith, C. D. (2004). Schizophrenia and theory of mind. *Psychological Medicine*, *34*(3), 385–389. https://doi.org/10.1017/s0033291703001326

Frith, C. D., & Singer, T. (2008). Review. The role of social cognition in decision making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3875–3886. https://doi.org/10.1098/rstb.2008.0156

Frith, C. (2004). Schizophrenia and theory of mind. *Psychological Medicine*, *34*, 385–389. https://doi.org/10.1017/S0033291703001236

Frith, Chris, & Frith, U. (2006). The Neural Basis of Mentalizing. *Neuron*, *50*(4), 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *37*(7), 1297–1310. https://doi.org/10.1016/j.neubiorev.2013.03.023

Gershman, S. J. (2015). A Unifying Probabilistic View of Associative Learning. *PLoS Computational Biology*, *11*(11), 1–20. https://doi.org/10.1371/journal.pcbi.1004567

Gradin, V. B., Pérez, A., MacFarlane, J. A., Cavin, I., Waiter, G., Engelmann, J., … Steele, J. D. (2015). Abnormal brain responses to social fairness in depression: An fMRI study using the Ultimatum Game. *Psychological Medicine*, *45*(6), 1241–1251. https://doi.org/10.1017/S0033291714002347

Gradin, Victoria B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., … Steele, J. D. (2011). Expected value and prediction error abnormalities in depression and schizophrenia. *Brain*, *134*(6), 1751–1764. https://doi.org/10.1093/brain/awr059

Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. *Nature Reviews Neuroscience*, *16*(10), 620–631. https://doi.org/10.1038/nrn4005

Greene, D. J., Colich, N., Iacoboni, M., Zaidel, E., Bookheimer, S. Y., & Dapretto, M. (2011). Atypical neural networks for social orienting in autism spectrum disorders. *NeuroImage*, *56*(1), 354–362. https://doi.org/10.1016/j.neuroimage.2011.02.031

Gromann, P. M., Heslenfeld, D. J., Fett, A. K., Joyce, D. W., Shergill, S. S., & Krabbendam, L. (2013). Trust versus paranoia: Abnormal response to social reward in psychotic illness. *Brain*, *136*(6), 1968–1975. https://doi.org/10.1093/brain/awt076

Gu, R., Huang, W., Camilleri, J., Xu, P., Wei, P., Eickhoff, S. B., & Feng, C. (2019). Love is analogous to money in human brain: Coordinate-based and functional connectivity metaanalyses of social and monetary reward anticipation. *Neuroscience and Biobehavioral Reviews*, *100*(January), 108–128. https://doi.org/10.1016/j.neubiorev.2019.02.017

Gunderson, J. G., Herpertz, S. C., Skodol, A. E., Torgersen, S., & Zanarini, M. C. (2018). Borderline personality disorder. *Nature Reviews Disease Primers*, *4*, 1–21. https://doi.org/10.1038/nrdp.2018.29

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233–1235. https://doi.org/10.1038/nn.4080

Haker, H., Schneebeli, M., & Stephan, K. E. (2016). Can Bayesian theories of autism spectrum disorder help improve clinical practice? *Frontiers in Psychiatry*, *7*(JUN), 1–17. https://doi.org/10.3389/fpsyt.2016.00107

Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154. https://doi.org/10.1007/BF02172093

Harlé, K. M., Guo, D., Zhang, S., Paulus, M. P., & Yu, A. J. (2017). Anhedonia and anxiety underlying depressive symptomatology have distinct effects on reward-based decision-making. *PLoS ONE*, *12*(10), 1–13. https://doi.org/10.1371/journal.pone.0186473

Henriques, J. B., & Davidson, R. J. (2000). Decreased responsiveness to reward in depression. *Cognition and Emotion*, *14*(5), 711–724. https://doi.org/10.1080/02699930050117684

Herpertz, S. C., & Bertsch, K. (2015). A New Perspective on the Pathophysiology of Borderline Personality Disorder: A Model of the Role of Oxytocin. *American Journal of Psychiatry*, *172*(9), 840–851. https://doi.org/10.1176/appi.ajp.2015.15020216

Huys, Q. J. M., Maia, T. V, & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., denOuden, H. E. M., & Stephan, K. E. (2013). Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron*, *80*(2), 519–530. https://doi.org/10.1016/j.neuron.2013.09.009

Jardri, R., Duverne, S., Litvinova, A. S., & Dene, S. (2017). *Experimental evidence for circular inference in schizophrenia*. https://doi.org/10.1038/ncomms14218

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *Npj Science of Learning*, *2*(1), 1–9. https://doi.org/10.1038/s41539-

017-0009-2

Juckel, G., Schlagenhauf, F., Koslowski, M., Wüstenberg, T., Villringer, A., Knutson, B., … Heinz, A. (2006). Dysfunction of ventral striatal reward prediction in schizophrenia. *NeuroImage*, *29*(2), 409–416. https://doi.org/10.1016/j.neuroimage.2005.07.051

Kapur, S. (2003). Psychosis as a State of Aberrant Salience: and Pharmacology in Schizophrenia. *Am J Psychiatry*, *160*, 13–23.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, R. (2005). Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science*, *308*(5718), 78–83. https://doi.org/10.1126/science.1108062

King-Casas, Brooks, Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science (New York, N.Y.)*, *321*(5890), 806–810. https://doi.org/10.1126/science.1156902

Klein-Flügge, M. C., Hunt, L. T., Bach, D. R., Dolan, R. J., & Behrens, T. E. J. (2011). Dissociable Reward and Timing Signals in Human Midbrain and Ventral Striatum. *Neuron*, *72*(4), 654–664. https://doi.org/10.1016/j.neuron.2011.08.024

Kronbichler, L., Stelzig-Schöler, R., Pearce, B. G., Tschernegg, M., Said-Yürekli, S., Crone, J. S., … Kronbichler, M. (2019). Reduced spontaneous perspective taking in schizophrenia. *Psychiatry Research - Neuroimaging*, *292*(August), 5–12. https://doi.org/10.1016/j.pscychresns.2019.08.007

Kupferberg, A., Bicks, L., & Hasler, G. (2016). Social functioning in major depressive disorder. *Neuroscience and Biobehavioral Reviews*, *69*, 313–332. https://doi.org/10.1016/j.neubiorev.2016.07.002

Langdon, R., Seymour, K., Williams, T., & Ward, P. B. (2017). Automatic attentional orienting to other people's gaze in schizophrenia. *Quarterly Journal of Experimental Psychology*, *70*(8), 1549–1558. https://doi.org/10.1080/17470218.2016.1192658

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, *8*(May), 1–10. https://doi.org/10.3389/fnhum.2014.00302

Lis, S., & Kirsch, P. (2016). Neuroeconomic Approaches in Mental Disorders. In M. Reuter & C. Montag (Eds.), *Neuroeconomics* (pp. 311–330). https://doi.org/10.1007/978-3-642-35923-1_16

Lockwood, P. L., & Klein-Flügge, M. (2019). *Computational modelling of social cognition and behaviour – a reinforcement learning primer*. 1–28.

Lomakina, E. I., Mathys, C. D., Daunizeau, J., Friston, K. J., Iglesias, S., Stephan, K. E., &

Brodersen, K. H. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*(November), 1–24. https://doi.org/10.3389/fnhum.2014.00825

Lowyck, B., Luyten, P., Vanwalleghem, D., Vermote, R., Mayes, L. C., & Crowley, M. J. (2015). What's in a face? Mentalizing in borderline personality disorder based on dynamically changing facial expressions. *Personality Disorders: Theory, Research, and Treatment*, *7*(1), 72–79. https://doi.org/10.1037/per0000144

Lynch, T. R., Rosenthal, M. Z., Kosson, D. S., Cheavens, J. S., Lejuez, C. W., & Blair, R. J. R. (2006). Heightened sensitivity to facial expressions of emotion in borderline personality disorder. *Emotion (Washington, D.C.)*, *6*(4), 647–655. https://doi.org/10.1037/1528-3542.6.4.647

Maia, T. V, & Frank, M. J. (2017). An Integrative Perspective on the Role of Dopamine in Schizophrenia. *Physiology & Behavior*, *176*(1), 139–148. https://doi.org/10.1016/j.physbeh.2017.03.040

Mathys, C. (2016). How could we get nosology from computation? *Computational Psychiatry: New Perspectives on Mental Illness*, *20*, 121–138. Retrieved from https://mitpress.mit.edu/books/computational-psychiatry

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*(November), 1–24. https://doi.org/10.3389/fnhum.2014.00825

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*(May), 1–20. https://doi.org/10.3389/fnhum.2011.00039

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

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*(7010), 760–767. https://doi.org/10.1038/nature03015

Mukherjee, D., Lee, S., Kazinka, R., D Satterthwaite, T., & Kable, J. W. (2020). Multiple Facets of Value-Based Decision Making in Major Depressive Disorder. *Scientific Reports*, *10*(1), 3415. https://doi.org/10.1038/s41598-020-60230-z

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology*, *14*(6), 769–776. https://doi.org/10.1016/j.conb.2004.10.016

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making. *Annual Review of Psychology*, *68*(1), 73–100. https://doi.org/10.1146/annurev-psych-010416-044216

Okruszek, Ł., Haman, M., Kalinowski, K., Talarowska, M., Becchio, C., Haman, M., & Manera, V. (2015). Impaired Recognition of Communicative Interactions from Biological Motion in Schizophrenia. *Plos One*, *10*(2), e0116793. https://doi.org/10.1371/journal.pone.0116793

Paulus, M. P., Huys, Q. J. M., & Maia, T. V. (2016). A Roadmap for the Development of Applied Computational PsychiatryModels for better diagnosis, prognosis and treatment. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1–22. https://doi.org/10.1016/j.bpsc.2016.05.001.A

Pfeiffer, U. J., Schilbach, L., Timmermans, B., Kuzmanovic, B., Georgescu, A. L., Bente, G., & Vogeley, K. (2014). Why we interact: On the functional role of the striatum in the subjective experience of social interaction. *NeuroImage*, *101*, 124–137. https://doi.org/10.1016/j.neuroimage.2014.06.061

Phan, K. L., Sripada, C. S., Angstadt, M., & McCabe, K. (2010). Reputation for reciprocity engages the brain reward center. *Proceedings of the National Academy of Sciences*, *107*(29), 13099–13104. https://doi.org/10.1073/pnas.1008137107

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(August), 596–600. https://doi.org/10.1126/science.aan3458

Pulcu, E., & Browning, M. (2017). Affective bias as a rational response to the statistics of rewards and punishments. *ELife*, *6*, 1–15. https://doi.org/10.7554/eLife.27879

Pulcu, E., & Browning, M. (2019). The Misestimation of Uncertainty in Affective Disorders. *Trends in Cognitive Sciences*, *23*(10), 865–875. https://doi.org/10.1016/j.tics.2019.07.007

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience*, *20*(8), 495–505. https://doi.org/10.1038/s41583-019-0179-4

Rescorla, R. A., & Wagner, A. R. (1972). *A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement*. 1–18. https://doi.org/10.1101/gr.110528.110

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A Neural Basis for Social Cooperation expertise, information, opportunities, and a host of ma. *Neuron*, *35*, 395–405. Retrieved from https://ac.els-cdn.com/S0896627302007559/1-s2.0-S0896627302007559-main.pdf?_tid=b30ec185-cd6d-4ad2-acd6-02d8be7580af&acdnat=1551065251_9b20b10ab6977b1373b83aaf75eaa9ae

Robson, S. E., Repetto, L., Gountouna, V. E., & Nicodemus, K. K. (2020). A review of neuroeconomic gameplay in psychiatric disorders. *Molecular Psychiatry*, *25*(1), 67–81. https://doi.org/10.1038/s41380-019-0405-5

Rothkirch, M., Tonn, J., Köhler, S., & Sterzer, P. (2017). Neural mechanisms of reinforcement learning in unmedicated patients with major depressive disorder. *Brain*, *140*(4), 1147–1157. https://doi.org/10.1093/brain/awx025

Roux, P., Smith, P., Passerieux, C., & Ramus, F. (2016). Preserved implicit mentalizing in schizophrenia despite poor explicit performance: Evidence from eye tracking. *Scientific Reports*, *6*, 1–9. https://doi.org/10.1038/srep34728

Ruff, C. C., & Fehr, E. (2014). *The neurobiology of rewards and values in social decision making*. (July). https://doi.org/10.1038/nrn3776

Safra, L., Chevallier, C., & Palminteri, S. (2019). Depressive symptoms are associated with blunted reward learning in social contexts. *PLoS Computational Biology*, *15*(7), 1–22. https://doi.org/10.1371/journal.pcbi.1007224

Saunders, K. E. A., Goodwin, G. M., & Rogers, R. D. (2015). Borderline personality disorder, but not euthymic bipolar disorder, is associated with a failure to sustain reciprocal cooperative behaviour: Implications for spectrum models of mood disorders. *Psychological Medicine*, *45*(8), 1591–1600. https://doi.org/10.1017/S0033291714002475

Schilbach, L. (2015). Eye to eye, face to face and brain to brain: Novel approaches to study the behavioral dynamics and neural mechanisms of social interactions. *Current Opinion in Behavioral Sciences*, *3*, 130–135. https://doi.org/10.1016/j.cobeha.2015.03.006

Schilbach, L. (2016). Towards a second-person neuropsychiatry. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686).

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Schilling, L., Wingenfeld, K., Löwe, B., Moritz, S., Terfehr, K., Köther, U., & Spitzer, C. (2012). Normal mind-reading capacity but higher response confidence in borderline personality disorder patients. *Psychiatry and Clinical Neurosciences*, *66*(4), 322–327. https://doi.org/10.1111/j.1440-1819.2012.02334.x

Schlagenhauf, F., Huys, Q. J. M., Deserno, L., Rapp, M. A., Beck, A., Heinze, H. J., … Heinz, A. (2014). Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *NeuroImage*, *89*, 171–180. https://doi.org/10.1016/j.neuroimage.2013.11.034

Schmack, K., Rothkirch, M., Priller, J., & Sterzer, P. (2017). Enhanced predictive signalling in schizophrenia. *Human Brain Mapping*, *38*(4), 1767–1779. https://doi.org/10.1002/hbm.23480

Schroeder, K., Fisher, H. L., & Schäfer, I. (2013). Psychotic symptoms in patients with borderline personality disorder. *Current Opinion in Psychiatry*, *26*(1), 113–119. https://doi.org/10.1097/yco.0b013e32835a2ae7

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

Schultz, Wolfram. (2007). Behavioral dopamine signals. *Trends in Neurosciences*, *30*(5), 203–210. https://doi.org/10.1016/j.tins.2007.03.007

Schultz, Wolfram. (2013). Updating dopamine reward signals. *Current Opinion in Neurobiology*, *23*(2), 229–238. https://doi.org/10.1016/j.conb.2012.11.012

Senju, A., & Johnson, M. H. (2009). The eye contact effect: mechanisms and development. *Trends in Cognitive Sciences*, *13*(3), 127–134. https://doi.org/10.1016/j.tics.2008.11.009

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind Eyes: An Absence of Spontaneous Theory of Mind in Asperger Syndrome*Science*, *325*(August), 883–885. https://doi.org/10.1126/science.1176170

Sevgi, M., Diaconescu, A. O., Henco, L., Tittgemeyer, M., & Schilbach, L. (2020). Social Bayes: Using Bayesian Modeling to Study Autistic Trait–Related Differences in Social Cognition. *Biological Psychiatry*, *87*(2), 185–193. https://doi.org/10.1016/j.biopsych.2019.09.032

Sharp, C. (2014). *The Social–Cognitive Basis of BPD: A Theory of Hypermentalizing BT - Handbook of Borderline Personality Disorder in Children and Adolescents* (C. Sharp & J. L. Tackett, eds.). https://doi.org/10.1007/978-1-4939-0591-1_15

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain Responses to the Acquired Moral Status of Faces. *Neuron*, *41*(4), 653–662. https://doi.org/10.1016/S0896-6273(04)00014-5

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

Stephan, K. E., & Mathys, C. (2014b). Computational approaches to psychiatry. *Current*

Opinion in Neurobiology, *25*, 85–92. https://doi.org/10.1016/j.conb.2013.12.007

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., … Corlett, P. R. (2018). The Predictive Coding Account of Psychosis. *Biological Psychiatry*, 1–10. https://doi.org/10.1016/j.biopsych.2018.05.015

Sterzer, P., Voss, M., Schlagenhauf, F., & Heinz, A. (2018). Decision-making in schizophrenia: A predictive-coding perspective. *NeuroImage*, (August 2017), 1–11. https://doi.org/10.1016/j.neuroimage.2018.05.074

Strauss, G. P., Waltz, J. A., & Gold, J. M. (2014). A review of reward processing and motivational impairment in schizophrenia. *Schizophrenia Bulletin*, *40*(SUPPL. 2), 107–116. https://doi.org/10.1093/schbul/sbt197

Stuke, H., Weilnhammer, V. A., Sterzer, P., & Schmack, K. (2019). Delusion Proneness is Linked to a Reduced Usage of Prior Beliefs in Perceptual Decisions. *Schizophrenia Bulletin*, *45*(1), 80–86. https://doi.org/10.1093/schbul/sbx189

Unoka, Z., Seres, I., Áspán, N., Bódi, N., & Kéri, S. (2009). Trust game reveals restricted interpersonal transactions in patients with borderline personality disorder. *Journal of Personality Disorders*, *23*(4), 399–409. https://doi.org/10.1521/pedi.2009.23.4.399

Walter, H., Ciaramidaro, A., Adenzato, M., Vasic, N., Ardito, R. B., Erk, S., & Bara, B. G. (2009). Dysfunction of the social brain in schizophrenia is modulated by intention type: An fMRI study. *Social Cognitive and Affective Neuroscience*, *4*(2), 166–176. https://doi.org/10.1093/scan/nsn047

Waltz, J. A., Schweitzer, J. B., Gold, J. M., Kurup, P. K., Ross, T. J., Jo Salmeron, B., … Stein, E. A. (2009). Patients with schizophrenia have a reduced neural response to both unpredictable and predictable primary reinforcers. *Neuropsychopharmacology*, *34*(6), 1567–1577. https://doi.org/10.1038/npp.2008.214

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

Weightman, M. J., Air, T. M., & Baune, B. T. (2014). A review of the role of social cognition in major depressive disorder. *Frontiers in Psychiatry*, *5*(NOV), 1–13. https://doi.org/10.3389/fpsyt.2014.00179

Winton-Brown, T. T., Fusar-Poli, P., Ungless, M. A., & Howes, O. D. (2014). Dopaminergic basis of salience dysregulation in psychosis. *Trends in Neurosciences*, *37*(2), 85–94. https://doi.org/10.1016/j.tins.2013.11.003

Wittmann, M. K., Lockwood, P. L., & Rushworth, M. F. S. (2018). Neural Mechanisms of Social Cognition in Primates. *Annual Review of Neuroscience*, *41*(1), 99–118.

https://doi.org/10.1146/annurev-neuro-080317-061450

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva.

Zandersen, M., & Parnas, J. (2019). Exploring schizophrenia spectrum psychopathology in borderline personality disorder. *European Archives of Psychiatry and Clinical Neuroscience*, (0123456789). https://doi.org/10.1007/s00406-019-01039-4

Acknowledgements

First and foremost, I would like to thank my supervisor Leonhard Schilbach for the opportunity to conduct my PhD within his research group. I thank Leo for all his support and encouragement in the projects, inspiring and thoughtful discussions and in particular for giving me the opportunity to realise my desire of conducting research with psychiatric patients. It was a pleasure to experience your enthusiasm and especially the ability to maintain the 'big picture', constantly asking the central question of how our research can eventually help patients.

I would also like to express my sincerest gratitude to Christoph Mathys for his generous and kind-hearted support and his patient and calm way of introducing me to computational modelling. In addition, I would like to thank Andreea Diaconescu and Meltem Sevgi for their valuable input in conducting my research and for doing substantial groundwork in the experimental design and computational modelling. I would also like to thank my TAC members Kathrin Koch and Christian Sorg for their valuable contributions and discussions as well as Philipp Sterzer for his immediate consent to review this thesis.

Moreover, I would like to thank all members of the social neurosciences lab for creating a great working atmosphere that I will definitely miss. I am especially indebted to Marie-Luise and Juha for their relentless support with data collection, analysis and great scientific advice throughout! I really don't know what I would have done without you <3. Imme and Laura, thank you for sticking around and also, making this challenging PhD time so much more bearable. Dimitris, thank you for your initial introduction to learning models, which helped a lot. I would also like to thank the GSN for offering an indispensable atmosphere for young researchers and great opportunities in attending conferences and social events. It has been a pleasure to be part of this PhD program!

Further thanks go to Max Schneider, Laura Leuchs, Victor Spoormaker, Michael Czisch and Philipp Sämann for inspiring journal clubs and helpful inputs to my research. A great thanks also goes to Ines Eidner for her patient help in data acquisition! Juha and Alec, thank you so much for proof-reading this thesis and your valuable comments. I would also like to thank all the participants that have taken part in my projects!

Finally, I want to thank my family and my partner Julius. I will refrain from heart-releasing expressions here. You know what you mean to me!

Declaration of Author Contribution

Bayesian modelling captures inter-individual differences in social belief computations in the putamen and insula. **Henco L**, Brandi M-L, Lahnakoski JM, Diaconescu AO, Mathys C & Schilbach L. 2020. Cortex

The author of this thesis is the first author of the publication; AOD and LS designed the research; **LH** and MLB implemented the experiment with the MR environment and collected the data; **LH** performed the computational data analysis with help of AOD, CM and the fMRI analysis with help of MLB and JML; **LH** wrote the manuscript; all authors reviewed and edited the manuscript. LS was the supervisor of the project and provided funding.

Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. **Henco L**, Diaconescu AO, Lahnakoski JM, Brandi M-L, Hörmann S, Hennings J, Hasan A, Papazova I, Strube W, Bolis D, Schilbach L & Mathys C. 2020. Submitted for publication

The author of this thesis is the first author of the publication; AOD and LS designed the research; JML programmed the online experiment; **LH**, SH, JH, AH, IP and WS recruited participants and collected data; **LH** performed the computational data analysis with help of CM under the supervision of LS; **LH** wrote the manuscript; all authors reviewed and edited the manuscript. LS was the supervisor of the project and provided funding.

Munich, 13.03.2020

Lara Henco

Munich, 13.03.2020

Prof. Dr. Leonhard Schilbach

List of Publications

Henco, L., Brandi, M. L., Lahnakoski, J., Diaconescu, A. O., Mathys, C & Schilbach, L. Bayesian modelling captures inter-individual differences in social belief computations in the putamen and insula (2020). *Cortex (*accepted for publication on 14th of February 2020).

Sevgi, M., Diaconescu, A. O., **Henco, L.,** Tittgemeyer, M., & Schilbach, L. (2020). Social Bayes: Using Bayesian Modeling to Study Autistic Trait–Related Differences in Social Cognition. *Biological Psychiatry*, *87*(2), 185-193.

Albantakis, L., Parpart, H., Krankenhagen, M., Böhm, J., **Henco, L.**, Brandi, M. L., & Schilbach, L. (2018). Autismus-Spektrum-Störungen (ASS) im Erwachsenenalter–Persönlichkeitsprofile und Begleiterkrankungen. *PTT-Persönlichkeitsstörungen: Theorie und Therapie*, *22*(1), 56-71.

Parpart, H., Krankenhagen, M., Albantakis, L., **Henco, L.**, Friess, E., & Schilbach, L. (2018). Schematherapie-informiertes, soziales Interaktionstraining. *Psychotherapeut*, *63*(3), 235-242.

Under review

Henco L., Diaconescu, A.O., Lahnakoski J.M., Brandi M.-L., Hörmann S., Hennings J., Hasan A., Papazova I., Strube W., Bolis, D., Schilbach L & Mathys C. (2020). Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder.

Albantakis, L., Brandi, M.-L., Zillekens, I. C., **Henco, L.**, Weindl, L., Thaler, H., Schliephake, L. M., Timmermans, B. & Schilbach, L. (2019). Autism and alexithymia – relevance for comorbid depression, social phobia, and psychosocial outcomes in adults with autism.

In preparation

Westenberg, E., Albantakis, L., **Henco, L.**, Lahnakoski, J. M., Schilbach, L., Brandi, M. L. (2020). Increased cognitive effort during explicit mentalizing in autism – A pupillometry study.

Eidesstattliche Versicherung/Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation „Using Bayesian modelling to uncover the behavioural and neural mechanisms of social learning and decision-making in healthy controls & psychiatric disorders“ selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation "Using Bayesian modelling to uncover the behavioural and neural mechanisms of social learning and decision-making in healthy controls & psychiatric disorders" is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

Munich, 13.03.2020

Lara Henco