CORTEX 131 (2020) 221–236

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special Issue "Understanding Others": Research Report

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,c,d], Andreea O. Diaconescu [e,f,g], Christoph Mathys [f,h,i,1] and Leonhard Schilbach [a,b,j,k,1]

[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] Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany
[d] Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[e] Department of Psychiatry (UPK), University of Basel, Basel, Switzerland
[f] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland
[g] Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), University of Toronto, Canada
[h] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
[i] Interacting Minds Centre, Aarhus University, Aarhus, Denmark
[j] Department of Psychiatry, Ludwig-Maximilians-Universität, Munich, Germany
[k] Outpatient Clinic and Day Clinic for Disorders of Social Interaction, Max Planck Institute of Psychiatry, Munich, Germany

ARTICLE INFO

Article history:
Received 30 May 2019
Reviewed 22 October 2019
Revised 21 December 2019

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

☆ 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, including presentation code and MATLAB analysis code has been made publicly accessible here: https://osf.io/keztf/.

* *Corresponding author*. Independent Max Planck Research Group for Social Neuroscience, Max Planck Institute of Psychiatry, Kraepelinstr. 8-10, 80804, Munich, Germany.
E-mail address: lara_henco@psych.mpg.de (L. Henco).
[1] The authors contributed equally to this work.

https://doi.org/10.1016/j.cortex.2020.02.024

Accepted 14 February 2020
Published online 25 April 2020

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

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.

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, 2020; 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, Doll, & Amodio, 2015; Lockwood, Apps, Valton, Viding, & Roiser, 2016; Lockwood & Klein-Flügge, 2020). 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 (Diaconescu et al., 2017; King-Casas et al., 2005). 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, Mathys, Brodersen, Kasper, Piccirelli, denOuden, 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 (Fig. 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, Mathys, Brodersen, Kasper, Piccirelli, denOuden, 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 (Fig. 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 (Fig. 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 sec 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 sec) before the outcome (correct/wrong) was presented for 2 sec. During choice, both cards were showing reward values (ranging from 1 to 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 sec) and 12 of these inter-trial intervals were jittered at longer durations (12–15 sec), 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 Fig. 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)

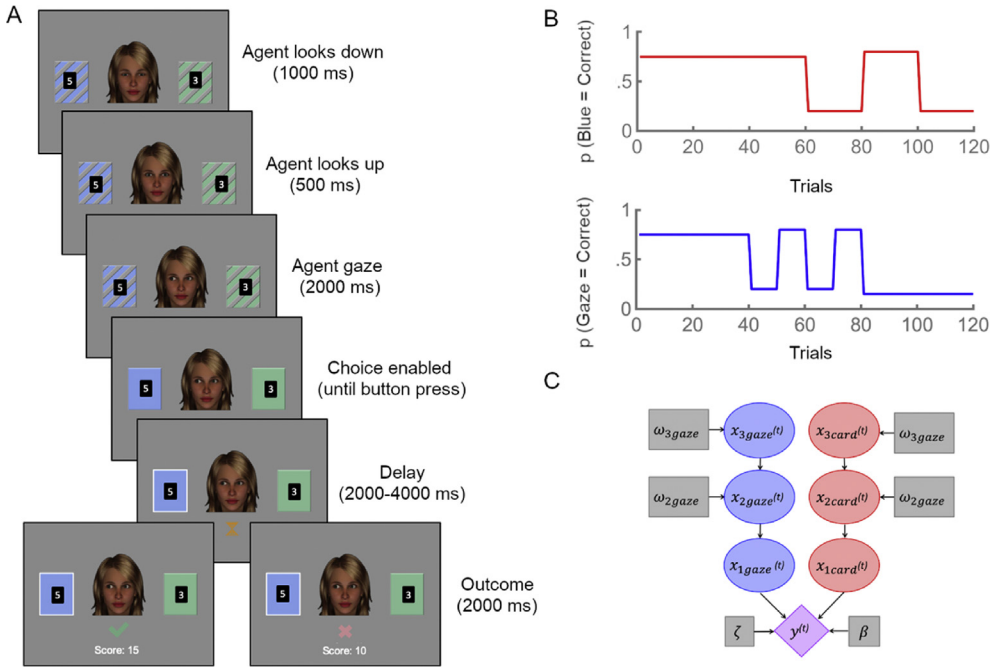

Fig. 1 – A: Trial flow and task design. On every trial, participants choose one of two cards (green & blue). After the choice is logged, an hourglass is presented followed by a green tick or a red cross depending on whether the response was correct or wrong. With every correct response, the score of 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).

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 1024 × 768 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, Mathys, Brodersen, Kasper, Piccirelli, denOuden, 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 (Fig. 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_{1card} 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.

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

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 2-4) 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.

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

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

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

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

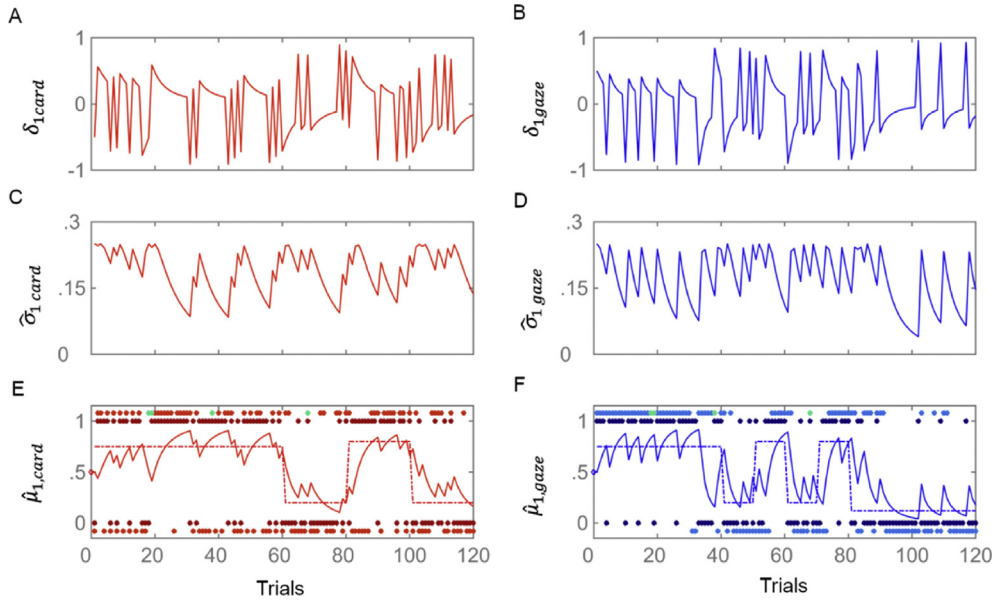

Fig. 2 – Example of participant-specific learning trajectories for both cues. A) Prediction error δ_{1card} (red) about trial outcome in terms of the non-social cue and B) prediction error δ_{1gaze}(blue) about the trial outcome in terms of the social cue. C) Variance (uncertainty) of prediction about non-social cue $\hat{\sigma}_{1card}$ and D) social cue $\hat{\sigma}_{1gaze}$. 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.

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 $\hat{\mu}^{(t)}_{1,gaze}$and card $\hat{\mu}^{(t)}_{1,card}$contingency with outcome (called 'accuracy' in what follows), weighted by precision was mapped onto decisions (Equation (5)). The combined belief was modelled as the sum of the posterior predictive expectation of gaze accuracy $\hat{\mu}^{(t)}_{1,gaze}$ and card accuracy $\hat{\mu}^{(t)}_{1,card}$ weighted by weights $w^{(t)}_{gaze}$ and $w^{(t)}_{card}$ (Equations (6) and (7)), 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 $\hat{\mu}^{(t)}_{1,card}$ was translated into gaze space, so that $\hat{\mu}^{(t)}_{1,card} = \hat{\mu}^{(t)}_{1,card}$ if the gaze went to the blue card, but $\hat{\mu}^{(t)}_{1,card} = 1 - \hat{\mu}^{(t)}_{1,card}$ if the gaze went to the green card. The precisions $\hat{\pi}_1$ (Equations (8) and (9)) 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 $\hat{\mu}^{(t)}_1$ moves away from .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 Fig. 3 illustrate the implications of high and low ζ values for decision-making.

$$b^{(t)} = w^{(t)}_{gaze}\hat{\mu}^{(t)}_{1,gaze} + w^{(t)}_{card}\hat{\mu}^{(t)}_{1,card} \tag{5}$$

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

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

$$\hat{\pi}^{(t)}_{1,gaze} = \frac{1}{\hat{\mu}^{(t)}_{1,gaze}\left(1 - \hat{\mu}^{(t)}_{1,gaze}\right)} \tag{8}$$

$$\hat{\pi}^{(t)}_{1,card} = \frac{1}{\hat{\mu}^{(t)}_{1,card}\left(1 - \hat{\mu}^{(t)}_{1,card}\right)} \tag{9}$$

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 $Prob^{(t)}_{gaze}$ was modelled as a logistic sigmoid (softmax) function of combined belief $b^{(t)}$ (Equation (5)), weighted by the expected reward of the card when following the advice r_{gaze}or not $r_{notgaze}$ (Equation (10)).

$$\begin{aligned} prob_{gaze} &= p(y^{(t)} = 1) \\ &= 1 \Big/ \left(1 + \exp\left(-\gamma^{(t)}\left(r^{(t)}_{gaze}b^{(t)} - r^{(t)}_{notgaze}\left(1 - b^{(t)}\right)\right)\right)\right) \end{aligned} \tag{10}$$

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

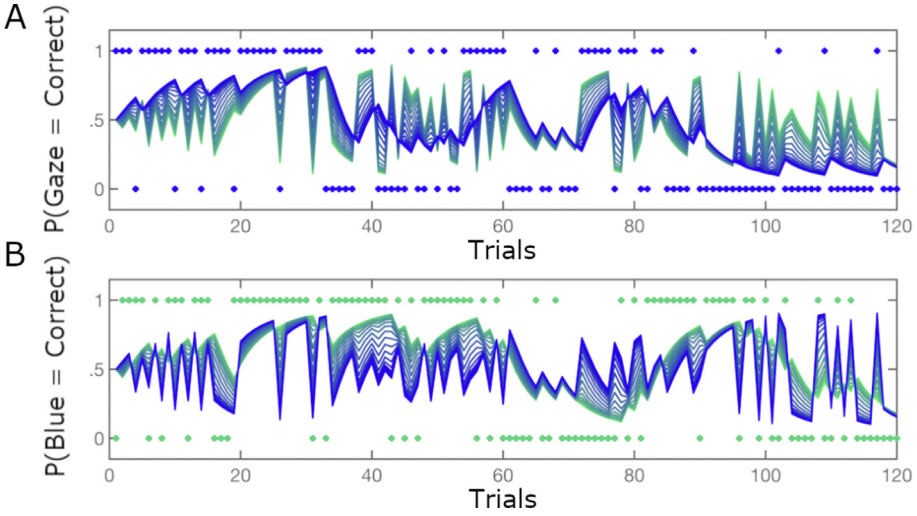

Fig. 3 – Simulation for an agent with same perceptual parameters but different social cue weighting. The plot shows A) the different probability trajectories for taking the advice (p (y = 1 | b) with varying ζ values (highest values (log(5) coded in blue, lowest values (log(-5) coded in light colours) 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).

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 (11)), (2) a combination of the log-volatility of the third level for gaze and participant-specific decision noise (Equation (12)), (3) a combination of the log-volatility of the third level for card and participant-specific decision noise (Equation (13)) or (4) the participant-specific decision noise alone (Equation (14)).

$$1)\ \gamma^{(t)} = \beta \exp\left(-\hat{\mu}^{(t)}_{3,card} - \hat{\mu}^{(t)}_{3,gaze}\right) \tag{11}$$

$$2)\ \gamma^{(t)} = \beta \exp\left(-\hat{\mu}^{(t)}_{3,gaze}\right) \tag{12}$$

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

$$4)\ \gamma^{(t)} = \beta \tag{14}$$

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 (14)). 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 (Rigoux, Stephan, Friston, & Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009).

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 a post-experimental questionnaire, asking participants how much they used the gaze (on a scale 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 .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-T 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 1 × 1 × 1 mm. Whole brain functional images were acquired (AC-PC-orientation, interleaved bottom-up, slice number = 40, inter-slice gap = .5 mm, TE = 20 msec, TR = 2000 msec, 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 × 2 × 2 mm for functional images and 1 × 1 × 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

2.7.1. 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 ($\hat{\mu}^{(t)}_{1,gaze}$), the variance (i.e., uncertainty) of this belief ($\hat{\sigma}^{(t)}_{1,gaze}$), and the variance about the probability of the winning card colour ($\hat{\sigma}^{(t)}_{1,card}$). We did not use $\hat{\mu}^{(t)}_{1,card}$ 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 $\delta^{(t)}_{1gaze}$ and the non-social prediction error signal $\delta^{(t)}_{1card}$ (an example of these trajectories can be seen in Fig. 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 $\hat{\mu}^{(t)}_{1,gaze}$. The outcome phase of the task (modelled for 2 sec starting at outcome presentation) was parametrically modulated by four regressors: The first regressor contained $\delta^{(t)}_{1gaze}$ neutralised, where the choice was wrong by setting the regressor's value to zero. In the second regressor, $\delta^{(t)}_{1gaze}$ was set to zero where the choice was correct. This way we could evaluate $\delta^{(t)}_{1gaze}$ 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 $|\delta^{(t)}_{1card}|$ 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 $\hat{\mu}^{(t)}_{1,gaze}$ and $\hat{\sigma}^{(t)}_{1,gaze}$, we estimated $\hat{\sigma}^{(t)}_{1,gaze}$ and $\hat{\sigma}^{(t)}_{1,card}$ in a separate GLM, which was the same as the one described above but differed in that the choice phase was modulated by $\hat{\sigma}^{(t)}_{1,gaze}$ and $\hat{\sigma}^{(t)}_{1,card}$ and not by $\hat{\mu}^{(t)}_{1,gaze}$. For completeness, we also estimated a GLM that included all parametric regressors ($\hat{\mu}^{(t)}_{1,gaze}$, $\hat{\sigma}^{(t)}_{1,gaze}$ and $\hat{\sigma}^{(t)}_{1,card}$) 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.

2.7.2. 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 ($\delta^{(t)}_{1gaze} > 0$, i.e., gaze helpful) and negative social prediction errors ($\delta^{(t)}_{1gaze} < 0$, 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 $\delta^{(t)}_{1gaze}$ 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 $\hat{\mu}^{(t)}_{1,gaze}$, we included the social weighting factor ζestimated 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 ζand 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 $-\hat{\sigma}^{(t)}_{1,card}$ 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 $p = .001$, followed by a cluster-level correction for multiple testing, with significance defined as cluster-level *p*-values < .05 after correction for family-wise error rate (FWE).

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 ($768 \times 1024 \times 120$) 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 $\widehat{\mu}_{3,card}$ and participant-specific decision noise β (XP = .937; PXP = .627; EXP_P = .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 (10)) 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 (Fig. 3A) and the non-social information (Fig. 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).

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 [$r_s(48) = .453$, $p = .001$] and the subjective report of finding the gaze helpful [$r_s(48) = .292$, $p = .044$] (Fig. 4A and B). The social weighting factor ζ was positively correlated with the proportion of trials in which the gaze was followed [$r_s(48) = .487$, $p < .001$] (Fig. 4C). The same was the case for the subjective report of using the gaze [$r_s(48) = .449$, $p = .001$]. Furthermore, when looking at advice-taking behaviour, the repeated measures ANOVA revealed a main effect of task phase [$F(2,96) = 57.050$, $p < .001$, $\eta^2 = .543$] 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 [$t(50) = 7.357$, $p < .001$, $d = 1.04$]. During the low-accuracy phase (20%) participants chose the advice significantly less compared to the volatile [$t(50) = 2.911$, $p = .016$, $d = .41$] and compared to the high accuracy phase [$t(50) = 8.340$, $p < .001$, $d = 1.179$]. There was a main effect of the covariate ζ [$F(1,48) = 17.54$, $p < .001$, $\eta^2 = .268$] and a significant interaction between the covariate ζ and the magnitude of the effect of task phase on behaviour [$F(2,96) = 6.832$, $p = .002$, $\eta^2 = .125$] 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 ($p < .001$, uncorrected) during choice phase were significantly correlated with ζ (Fig. 4D).

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	.464	.098	.077	.031	.289	.042
PXP	.627	.067	.067	.067	.105	.067
XP	.937	0	0	0	.063	0

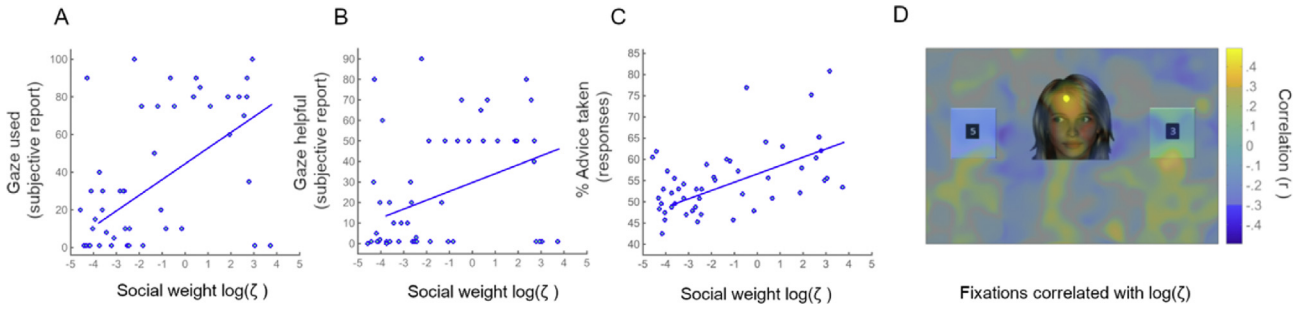

Fig. 4 – A) Association between estimated values of computational parameter ζ and subjective reports of having used the gaze and B) finding it helpful during decision-making. C) Association between ζ and % of trials where gaze was followed. D) Mean proportion of fixations on face area during all trials and ζ; Pixel-wise analysis of smoothed fixation data revealed that ζ is correlated with the time people spend looking at the face ($p < .001$, uncorrected) during the choice phase of the trials.

3.4. *fMRI results*

3.4.1. *Social and non-social prediction and precision during decision-making*

During the choice phase of the task, the subjective predicted advice accuracy $\hat{\mu}^{(t)}_{1,gaze}$ 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 (Fig. 5 and Table 2). Self-reports of having used the gaze during decision-making were associated with higher activity related to $\hat{\mu}^{(t)}_{1,gaze}$in the right rectal gyrus, right and left putamen and insula (Fig. 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 A4). Significant clusters were neither found for the correlation with $1-\hat{\mu}^{(t)}_{1,gaze}$ (the subjective predicted probability of a misleading gaze) nor for the variance of the prediction $\hat{\sigma}^{(t)}_{1,gaze}$ and $1-\hat{\sigma}^{(t)}_{1,gaze}$. Results for $\hat{\mu}^{(t)}_{1,gaze}$ when estimated together with $\hat{\sigma}^{(t)}_{1,gaze}$ and $\hat{\sigma}^{(t)}_{1,card}$ in one GLM can be seen in Table A5 & A6.

In the choice phase of the task, the negative contrast on the variance of the belief about the winning card colour ($1-\hat{\sigma}^{(t)}_{1,card}$) 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$1-\hat{\sigma}^{(t)}_{1,card}$. No significant clusters were found for the positive contrast ($\hat{\sigma}^{(t)}_{1,card}$).

3.4.2. *Social and non-social prediction error during outcome*

For negative social prediction errors ($\delta^{(t)}_{1gaze}<0$, 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 ($\delta^{(t)}_{1gaze}>0$, 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 as when evaluating negative social prediction errors against baseline during incorrect outcomes. When we directly compared positive prediction error signals during

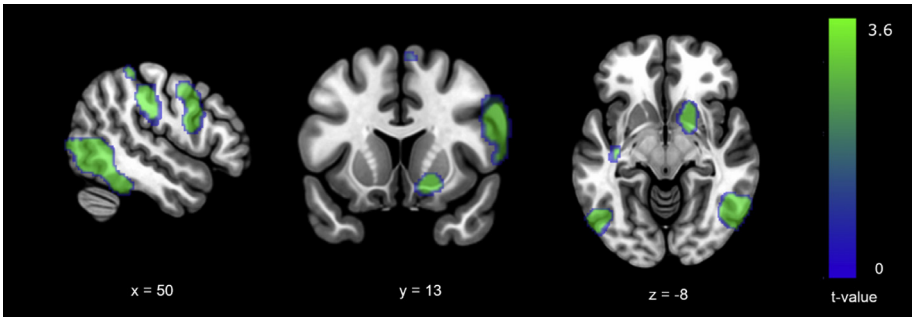

Fig. 5 – fMRI results for predicted accuracy of advice ($\hat{\mu}^{(t)}_{1,gaze}$) during the choice phase of the task. Cluster-forming threshold: $p < .001$, cluster-level threshold $p < .05$, FWE corrected. [x y z] coordinates refer to the MNI coordinates of the respective slices. See Table 2 for further information on cluster extents and peak voxel coordinates.

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

Region (left/right)	Pcluster	Cluster		MNI coordinates		
		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	.001	524	5.77	−50	−68	−8
L Middle Temporal Gyrus			3.29	−56	−58	2
White Matter	.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	.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 Cerebellum (VI)			3.93	−32	−40	−28

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.

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.

Region (left/right)	Pcluster	Cluster		MNI coordinates		
		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

Next, we looked at the absolute prediction error of the advice ($\delta^{(t)}_{1card}$), 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 $\delta^{(t)}_{1card}$ 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 $\delta^{(t)}_{1card}$. The results for the negative contrast on $\delta^{(t)}_{1card}$ looking at activity correlated with a decrease in surprise about the winning card colour can be seen in Table A. 7.

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

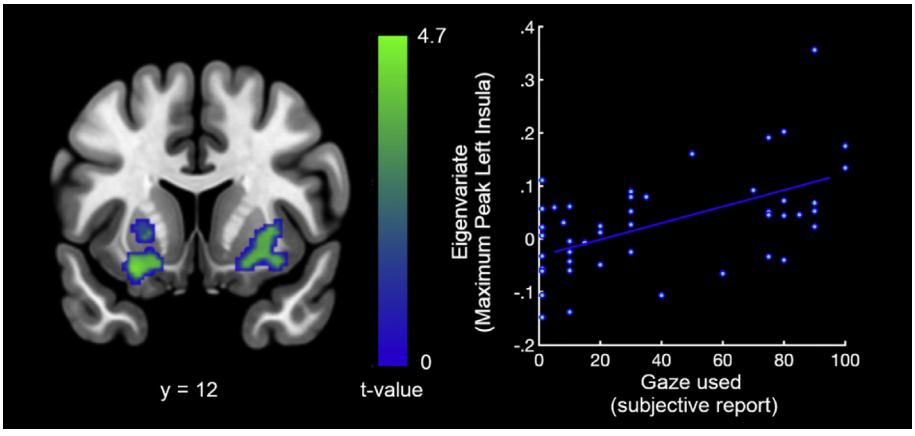

Fig. 6 – The contrast in the left shows brain areas showing differential responses to $\hat{\mu}^{(t)}_{1.gaze}$ as a function of the subjective report of having used the gaze during decision-making. The right plot depicts the correlation between the subjective report and the highest peak in the insula. Cluster-forming threshold: $p < .001$, cluster-level threshold $p < .05$, 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.

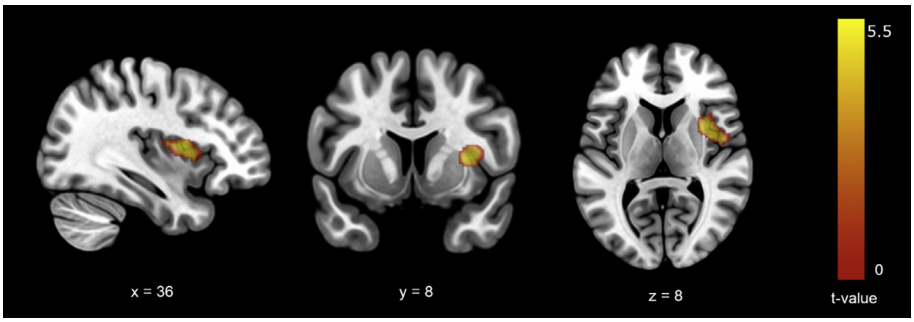

Fig. 7 – Significant clusters for the negative contrast of the variance of the prediction of the winning card colour ($\hat{\sigma}^{(t)}_{1.card}$) during the choice phase of the task. Cluster-forming threshold: $p < .001$ uncorrected, cluster-level threshold $p < .05$, FWE corrected. [x y z] coordinates refer to the MNI coordinates of the respective slices. See Table 4 for further information on cluster extents and peak voxel coordinates.

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

Region (left/right)	Pcluster	Cluster		MNI coordinates		
		k	Tpeak	x	y	z
R Insula Lobe	.003	381	5.51	36	6	10
R Rolandic Operculum			4.68	46	−2	14

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 ($\delta^{(t)}_{1gaze} < 0$, 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

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

Region (left/right)	Pcluster	Cluster		MNI coordinates		
		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 Gyrus	.009	312	4.87	0	4	64

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

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

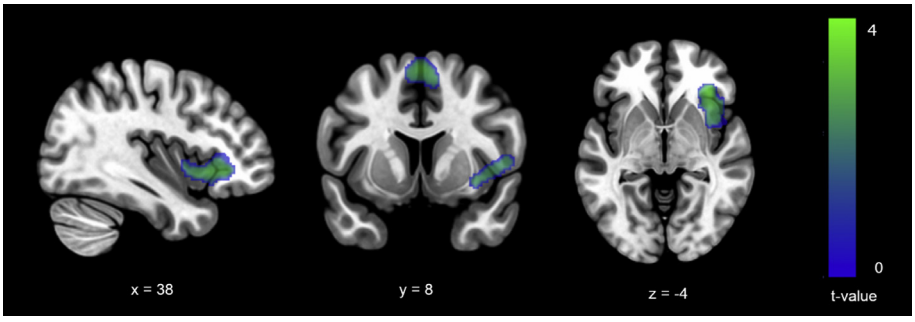

Fig. 8 – Neural correlates of ($\delta^{(t)}_{1gaze}$) during wrong choice outcomes. The negative contrast on the parametric modulator of the outcome phase reflects BOLD activity in regions correlated with negative social prediction errors. Cluster-forming threshold: $p < .001$, cluster-level threshold $p < .05$, 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.

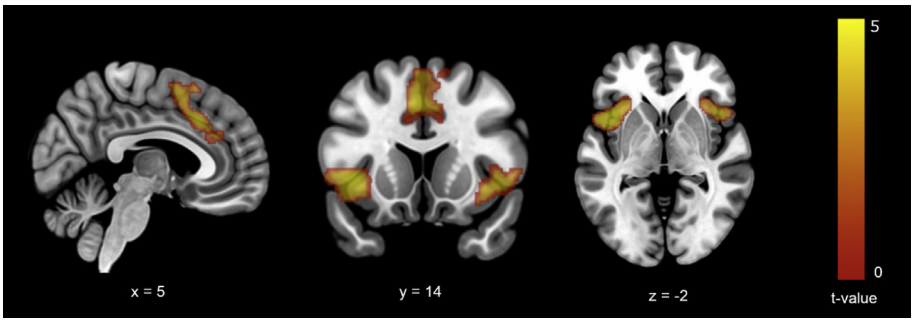

Fig. 9 – Neural correlates of absolute learning signal ($\delta^{(t)}_{1card}$). The positive contrast on the parametric modulator reflects BOLD activity in regions correlated with amount of surprise about the accuracy of the card colour when social cue was correct. Cluster-forming threshold: $p < .001$, cluster-level threshold $p < .05$, 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.

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 et al., 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 $\delta^{(t)}_{1card}$ in the left and right insula, a pattern resembling prediction error activation in a sensory learning paradigm (Iglesias, Mathys, Brodersen, Kasper, Piccirelli, denOuden, 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, Rieskamp, Krugel, & Heekeren, 2011; Delgado, Frank, & Phelps, 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, and 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., 2016, 2015; 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

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

Region (left/right)	Pcluster	Cluster		MNI coordinates		
		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 Gyrus			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)	.002	422	5.32	34	24	−8
R Insula Lobe			5.16	42	18	−4
All outcomes						
L Insula Lobe	.032	225	5.11	−44	12	−4
L Posterior-Medial Frontal Gyrus	.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

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 ($\hat{\mu}^{(t)}_{1,gaze}$), 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 overattribute the meaning and salience of social signals (Diaconescu et al., 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, cf. appendix Tab A.5). 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.

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.

Author contribution

Lara Henco: Conceptualization, Methodology, Investigation, Formal Analysis, Data Curation, Visualization, Writing – Original Draft, Project administration, Writing - review & editing.

Marie-Luise Brandi: Investigation, Formal Analysis, Writing – Review & Editing.

Juha Lahnakoski: Formal Analysis, Writing – Review & Editing, Visualization.

Andreea Oliviana Diaconescu: Methodology, Formal Analysis, Software, Validation, Writing – Review & Editing.

Christoph Mathys: Methodology, Formal Analysis, Software, Validation, Writing – Review & Editing, Visualization, Supervision.

Leonhard Schilbach: Methodology, Conceptualization, Formal analysis, Supervision, Funding Acquisition, Project Administration, Writing – Review & Editing.

Open practices

The study in this article earned an Open Materials badge for transparent practices. All digital materials associated with this experiment, presentation code, and analysis scripts can be https://osf.io/keztf/.

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.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2020.02.024.

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. (2001a). 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. (2001b). 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 and 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., et al. (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.

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.

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., et al. (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.
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., et al. (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., 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.
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.
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., et al. (2014). Synchronous brain activity across individuals underlies shared psychological perspectives. *Neuroimage, 100*, 316–324. https://doi.org/10.1016/j.neuroimage.2014.06.022.
Liebowitz, M. R. (1987). Social phobia. *Modern Problems in Pharmacopsychiatry, 22*(1987), 141–173.
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-Flügge, M. (2020). Computational modelling of social cognition and behaviour - a reinforcement learning primer. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsaa040.
Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience, 5*, 1–20. https://doi.org/10.3389/fnhum.2011.00039.
Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., et al. (2014). Uncertainty in perception and the hierarchical Gaussian filter. *Frontiers in Human Neuroscience, 8*(825). https://doi.org/10.3389/fnhum.2014.00825.
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* (pp. 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. (2008). 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., et al. (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. *Science, 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* (pp. 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.