Title

Neural computations underpinning the strategic management of influence in advice giving

Uri Hertz [1,2], Stefano Palminteri [3,4], Silvia Brunetti [1], Cecilie Olesen [1], Chris D Frith [5,6], Bahador Bahrami [1]

1 UCL Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, UK

2 School of Advanced Studies, University of London, Senate House, Malet Street, London WC1E 7HU, UK

3 Laboratore de Neurosciences Cognitives, Institut National de la Santé et de la Recherche Médicale, 75005, Paris, France.

4 Departement d'Études Cognitives, École Normale Supérieure, 75005, Paris, France.

5 Wellcome Trust Centre for Neuroimaging, University College London, 12 Queen Square, London WC1N 3BG, UK

6 Institute of Philosophy, School of Advanced Studies, University of London, Senate House, Malet Street, London WC1E 7HU, UK

Corresponding Author:

Uri Hertz

UCL Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, UK, Phone: +44-20-76795429, Email: u.hertz@ucl.ac.uk

Keywords:

Social Brain; Decision Making; Social Influence; Brain Valuation system; Ventral Striatum; rTPJ; mPFC; advice giving ; fMRI;

Abstract

Research on social influence has mainly focused on the target of influence (e.g. consumer, voter), while the cognitive and neurobiological underpinnings of the source of the influence (e.g. marketer, politician) remain unexplored. Here we show that advisers managed their influence over their client strategically, by modulating the confidence of their advice, depending on the interaction between adviser's level of influence on the client (i.e., being ignored or chosen by the client), and relative merit (i.e., their accuracy in comparison with a rival). Functional magnetic resonance imaging showed that these sources of social information were tracked in distinct regions of the brain's social and valuation systems: relative merit in the medial-prefrontal cortex, and selection by client in the temporo-parietal junction. In addition, increased functional connectivity between these two regions during positive merit trials predicted their effect on behaviour. Both sources of social information modulated the activity in the ventral striatum. These results further our understanding of human interactions and provide a framework for investigating the neurobiology of how we try to influence others.

Main Text

Introduction

It is hard to overstate the role of social influence in our lives. From the outcomes of political campaigns to our stand on issues such as global warming, immigration and taxation, the fate of the most critical matters in our public lives depends on persuaders working to influence public opinion[1]. Social influence is also intertwined with our everyday life, as we try to persuade our children, influence our bosses, and gain popularity among our friends. Research on social influence has been dominated by the motivation to understand the minds of the targets of influence (e.g. consumers, voters) in order to exert even more influence on them[2]. Far less is known about the cognitive and neurobiological processes at play in the persuader's mind. Here we ask what happens in the persuader's mind (e.g., spin doctors or financial advisers) when engaged in the process of influencing others.

Bayarri and DeGroot[3] proposed an advising strategy for the way influence-minded advisers should offer their predictions. They assumed that clients are affected by advisers' accuracy and confidence[4–7], i.e., they will be more likely to follow advisers that express confidence only when it is warranted, and discredit highly confident but inaccurate advisers. In this case, in order to be selected by the client, advisers should adapt their advice confidence deviance, i.e., the difference between expressed confidence and objective evidence, depending on their current influence on the client, between: (a) express radical advice when the adviser has low influence over a client. In other words, when the client is more likely to ignore the advisor, s/he should express higher confidence than permitted by objective evidence; (b) when the adviser aims to maintain and protect already high influence, s/he should sit on the fence with cautious, nuanced, advice. We will call this a 'competitive' strategy for acquiring and maintaining influence. To follow this strategy, the adviser should track his/her current level of influence on the client, which raises some predictions regarding the potential underlying neural mechanisms of influence management. First, we hypothesise that evaluating what others think about us is tracked by the social brain system[8–12], including the medial prefrontal cortex (mPFC) and the temporo-parietal junction (TPJ). In addition, we hypothesise that having one's advice preferred over that of a rival adviser activates the human brain's valuation network including the Ventral Striatum (VS)[13]. We therefore predict that the human brain's social and valuation mechanisms constitute the neuronal substrates of the influence minded, competitive strategy of advice giving.

Social rank theory[14,15] proposes an alternative account of advising behaviour by positing that people are not motivated just by the desire to influence others' choices, but also by the fear of being excluded by the group. Consequently, the theory suggests, people may response according to their rank in a group. Lower rank individuals therefore adopt submissive behaviours such as low expressions of confidence, backing down and eye gaze avoidance[14]. Social rank theory suggests that, in contrast to the 'competitive' strategy described above, humans may adopt a 'defensive' strategy to manage influence by giving cautious advice when they are ignored by the client (i.e. when their influence is low) and exaggerated advice when their influence is high. Although the 'defensive' strategy gives the opposite behavioural prediction compared to the 'competitive' strategy, it also involves adjusting advice confidence according to the current level of influence, and therefore leads to a similar set of predictions about the neuronal substrates of exerting influence.

Social rank theory also underscores the importance of an active process of social comparison, by which an adviser can evaluate his/her rank by tracking his/her performance relative to rivals [16–20]. In this view, relative performance or merit may also affect advising behaviour. Evaluating one's rank through social comparison implies evaluation of others' performance and comparing it with that of oneself, a computation possibly underpinned by the mPFC [13,21–25] and the valuation system[26–29].

Here we set out to examine how people strategically manage their influence, and tested the hypotheses we derived above regarding the neural computations underlying this process. First, we examined the behavioural predictions of a normative model of advice giving [3] against those drawn from social rank theory [14]. We asked if people would give over confident ('competitive' strategy [3]) or under confident ('defensive' strategy [14,15]) advice when they are ignored by their client, and whether social comparison with a rival adviser plays a role in advising behaviour. We then sought to identify the neural mechanisms underlying the participants' attempts to influence others by advice giving, whether (and how) the social brain network and the valuation network track client appraisal and perform social comparison during strategic advice giving task.

Results

Behavioural Task

We devised a social influence scenario in which two advisers competed for influence over a client (Figure 1, online demo: http://www.urihertz.net/AdviserDemo). On a series of trials, a client is looking for a reward hidden in a black or a white urn. He relies on two advisers who have access to evidence about the probability of the reward being in the black or white urn. In the beginning of each trial (appraisal stage), the client chooses the adviser whose advice (given later) will determine which urn the client will open. The client's choice of adviser is displayed to the advisers, who then proceed to the evidence stage. They see a grid of black and white squares for half a second. The ratio between the black and white squares indicates the probability of the reward location. Next step is the advice stage. Each adviser declares his/her advice about the reward location using a 10 levels confidence scale, ranging from 'certainly in the black urn' (5B) to 'certainly in the white urn' (5W). Subsequently, both advice are then shown to both advisers and the client (showdown stage). Finally, at the outcome stage, the urn indicated by the chosen adviser's advice is opened, and its content is revealed to everyone. The next trial begins with the client selecting an adviser – the client can decide to switch advisers or to stick with the same adviser.

We were interested in the way advisers use advice confidence as a persuasive signal to manage their influence over the client. In our first experiment therefore, participants all played the role of *adviser*, while the rival adviser and the client's behaviour were governed by algorithms adapted from Bayarri and DeGroot[3] (see Methods). We examined participants in three cohorts: online (N=58), in the lab (N=29) and in-the-scanner (N=19). There were some minor differences between the groups, as online participants performed less trials than lab and in-the-scanner participants, and were paid a bonus for the number of times the client picked them while lab based and in-the-scanner participants received a fixed monetary compensation (see full details in Methods section). However, there were no differences in the task itself between cohorts. Finally, to examine whether the observed behaviour can be generalised to real life social interactions, we ran a lab based interactive experiment (48 participants organized in 16 groups of 3 people, thus comprising 32 advisers and 16 clients) in which all roles were played by participants.

The effect of selection by client and relative merit on advising

To assess behaviour, we examined trial-by-trial deviance of advice confidence from probabilistic evidence (Figure 2A). In zero-deviance policy confidence exactly matches the ratio of black to white squares in the evidence grid (Figure 1B), with confidence '5B' indicating close to 100% black squares in the grid, i.e. close to 100% probability of the coin being in the black urn, and confidence '1B' indicating evidence close to 50% probability of the coin being in the black urn. Advice would be deviant if the confidence is higher (positive deviance) or lower (negative deviance) than what would be expected given the probability indicated by the evidence. Participants' advice deviance was significantly higher than 0 ($t(105) = 16.8$, $p < 0.00001$), as they displayed systematic overconfidence in their advice. Importantly, participants' advice deviance was larger, i.e. more overconfident, under low influence (i.e. when ignored by the client) compared to high influence trials (paired t-test, $t(105) = 3.45$, $p = 0.001$) (Figure 2B). Moreover, participants adjusted their advising policy dynamically, increasing their confidence in periods when they received negative appraisals (not chosen by client) and reducing it during periods of high influence (Figure 2C).

This finding demonstrated how being selected over a rival shapes the attempt to influence others. However, our analysis so far focused only on selection by the client, i.e. current level of influence on the client, as the force shaping our attempt to influence others, while ignoring the effect of self-comparison with the rival adviser on advice giving, implicated by social rank theory [16–20]. A deeper understanding of the participant's advising strategy would require factoring in the interaction between the (exogenous) selection by client and the (endogenous) relative merit. To disentangle the impact of selection and relative merit on behaviour we employed a computational approach.

While trial-by-trial selection by client was explicitly available for the participants and for analysis, trial-by-trial variations of relative merit had to be inferred to examine its effect on advice giving. To this end we followed the model fitting approach used in behavioural and neuroimaging studies, to estimate latent subjective processes such as adviser's reliability, reward prediction errors and social comparisons [12,21,30]. We devised several models that made different assumptions regarding the way relative merit is tracked and affect advice deviance, and chose the winning model's estimate of trial-by-trial relative merit to determine how it affected behaviour. We noted that participants could evaluate the prognostic value of their own and their rival's advice every time the outcome (i.e. the content of the chosen urn) was revealed (Figure 3A). For example, strong advice for black urn (i.e. 5B) would have better prognostic value than a cautious advice supporting the same choice (i.e. 1B) if the black urn turned out to have a coin. Conversely, advising 1W would have better prognostic value than 5W if the white urn was chosen but turned out to be empty. Prognostic value is calculated by multiplying advice confidence with its accuracy (correct = 1, wrong=-1) (i.e. whether the urn suggested by the advice contained a coin, see methods for further details). On a given trial *t*, we operationalised the change in relative merit as the difference between the prognostic value of the participant's and the rival's advice such that:

$$[1]\ \Delta PrognosticValue(t) = Confidence_{You}(t) \cdot Accuracy_{You}(t) - Confidence_{Rival}(t) \cdot Accuracy_{Rival}(t)$$

$$[2] \qquad RelativeMerit(t+1) = (1-\gamma) \cdot RelativeMerit(t) + \gamma \cdot \Delta PrognosticValue(t)$$

In equation [2] γ is the aggregation rate of change to relative merit. Relative merit is therefore positive if the participant's advice had consistently higher prognostic value compared to rival's advice. Importantly, relative merit is calculated independently from selection by the client and gives a quantitative estimate of the latent subjective process of social comparison.

To quantify if and how this measure of relative merit could explain behaviour, we fitted a set of hierarchically nested computational models to advice deviance. Our most simple model had only one free parameter for systematic bias in advising, i.e. trait overconfidence and under-confidence (Bias Model). Our next model included the trial-by-trial selection by the client as well (Client Model). Positive values of the weight assigned to selection by client, would indicate that the participant followed a 'defensive' strategy, expressing higher confidence when selected by the client. Conversely, negative values for this parameter would correspond to the 'competitive' strategy. These models should account for the effects observed in our initial analysis, and therefore can serve as a baseline to compare more elaborate models that also include a relative merit measure. Our next model included the sign of the relative merit from equation [2] (Mixed model), and another model included of the above and the interaction between the selection by client and the relative merit (Interaction model, Equation [3]). We also fitted models that used the sign and amplitude of relative merit (see Supplementary Materials).

$$\begin{aligned}[3]\quad & AdviceDeviance_{Interaction}(t) = Bias + \beta_{Selection} \cdot Selection(t) + \ldots \\ & \beta_{Merit} \cdot sign\left(RelativeMerit(t)\right) + \beta_{Interaction} \cdot Selection(t) \cdot sign\left(RelativeMerit(t)\right)\end{aligned}$$

After fitting all models to the advice deviance data and compensating for the number of free parameters (see Supplementary table ST1, Supplementary Figure S2 and S3), we found that the interaction model (Equation [3]) gave the best fit to the empirical data. Our model fitting procedure estimated individual parameters for bias, γ , $\beta_{Selection}$, β_{Merit} and $\beta_{Interaction}$ (Supplementary Figure S3, Supplementary table ST1). Bias parameter was significantly higher than zero across participants, verifying our observation that participants were generally overconfident (Mean ± SME Beta: 0.74 ± 0.05, t(105) = 14.2, p < 10^{-10}) . In addition, the selection parameter ($\beta_{Selection}$) was significantly lower than zero across participants (Mean ± SME: -0.08± 0.02, t(105) = -3.63, p = 0.0004), supporting our direct comparison in Figure 2B, favouring the 'competitive' over 'defensive' hypothesis. While this parameter was significantly lower than 0, we observed high degrees of variability among individual, with a positive selection parameter estimated for one third of our participants (38) (Figure 2D). Previous works on social rank theory [14,31] suggested that negative self-perception, i.e. seeing self as inferior to others and less desirable, may lead to display of low confidence in social interaction and a greater receptivity to negative social signals. We reasoned that population variability in behavioural response to social selection and/or exclusion, as quantified here by our selection parameter may be accounted for by the individual differences in receptivity to negative social feedback. To test this hypothesis, we went back to our pool of participant and recruited N=63 of them to complete the Fear of Negative Evaluation questionnaire (FNE) [14,32]. We found that FNE score correlated with the participants' model estimate of selection parameter, as participants with higher FNE scored, i.e., more negative self-perception, were more likely to follow the defensive strategy (N=63, R = 0.26, R^2 = 0.067 p = 0.037). By linking previous research on social rank and our participants' behaviour, this latter finding provided evidence of external validity for our computational model-based results.

Estimated model parameters associated with relative merit, and the interaction term between relative merit and selection by client, were harder to interpret directly. Even greater individual differences were observed for these parameters. When averaged across participants, neither parameter was significantly different from zero (T(105)<0.5, P>0.6), with individuals varying in the effect of relative merit and interaction (see Table ST1). However, model selection and comparison showed beyond doubt that both parameters did provide the model with better power to capture behaviour. To examine the contribution of these parameters to explaining behaviour, we used the sign of the relative merit estimated by the interaction model using the fitted merit aggregation rate γ (Equation [2]), to label the trials as 'positive' vs 'negative' relative merit. Each trial was also categorised according to the selection by the client as 'ignored' or 'chosen'. We then examined advice deviance across the resulting 2x2 combinations of relative merit (positive vs. negative) and selection (selected vs. ignored). We found a significant interaction effect on advice deviance, as model fitting procedure suggested (Figure 3B, repeated measures ANOVA, F(1, 316) = 11.7, p = 0.001, see Figure S5 for breaking of these results across cohorts). Importantly, this analysis elucidated the nature of the interaction by showing that participants expressed significantly greater confidence in their advice on trials in which they assumed they had done better than their rival (i.e. positive relative merit) but nonetheless had been (perhaps unexpectedly) ignored by the client. To put it metaphorically, advisers shouted most loudly when they had reason to believe that their merits had been overlooked.

This behaviour, one may argue, can arise in response to the specific manner in which our algorithms, following Bayarri and DeGroot's assumptions[3], controlled the client and rival adviser behaviour. To confirm the validity and generality of our results, we ran a fully interactive experiment in which participants played the role of both advisers and the client (see Methods for details) and no agent's behaviour was under experimenter control. The scenario was the same as before, but now involved three participants engaged in a multiplayer game on played on three computers in three adjacent cubicles connected via the internet. We applied our analysis, i.e. model fitting and estimation of relative merit and selection by client effects, to the advisers' behaviour in this fully interactive experiment. The results replicated the main experiment with virtual agents: advice deviance was highest when the adviser felt that s/he was unjustly ignored, i.e. when their relative merit was positive but the client had ignored them (Figure 3C). Using a mixed effects ANOVA we found a significant interaction effect (F(1,96) = 5.05, p = 0.03), and a significant effect of relative merit (F(1,96) = 4.43 , p = 0.04).

To further examine the relation between the virtual and live experiments, we fitted data from all experimental cohorts, with a mixed effects ANOVA, with Relative Merit (positive/negative), Selection by Client (chosen/ignored) and Agents (live/virtual) as main factors, and subjects as random effect factor nested within the Agents factor. We did not observe any effect of Agents on advice, neither main effects nor interaction. Overall we observed a significant main effect of Relative Merit (F(1,412)

= 8.,5 p = 0.004), a significant main effect of Selection by Client (F(1,412) = 5.17, p = 0.024) and a significant interaction between the two factors (F(1,412) = 7.24, p = 0.008).

The results from the fully interactive experiment provided an additional replication of our main results, demonstrating the robustness of the findings and the ecological validity of the paradigm in generalizing the findings to interactive human behaviour.

Neural Correlates of Selection by Client and Relative Merit

Having established the effects of selection by client and relative merit on advising behaviour, we used the same paradigm to examine the neural correlates underlying the attempt to influence others. Employing functional magnetic resonance imaging (fMRI), each of the five stages of a trial was treated as an event in a fast event-related design.

At the outcome stage, participants had all the information necessary to calculate their relative merit (i.e., compare the prognostic value of their own and the rival's advice) before being appraised by the client (Figure 1B). Using a whole brain parametric modulations analysis, we found that activity in medial prefrontal cortex (mPFC)($p < 0.005$, FWE cluster size corrected $p < 0.05$, see Table ST2) tracked the trial-by-trial changes in relative merit, i.e. prognostic value comparisons with rival (Equation 1) (Figure 4A). Given the previously documented involvement of mPFC in evaluating and inferring other agents' traits such as reliability [12] and accuracy [23,24], and its role in mentalizing about others' intentions and beliefs [33], and evaluating one's rank in relation to others[22], the findings reported here support the role of mPFC in social comparison with rival adviser, tracking trial by trial relative performance used to update relative merit.

Using the computational model-based analysis described above, we labelled trials according to their relative merit (positive vs. negative), in addition to the selection by client labels (ignored vs. chosen). The resulting 2x2 combination of conditions allowed us to examine the changes in the blood oxygenation level dependent (BOLD) signal that correlated with the main effects of selection and its interaction with relative merit. Activity in the right temporo-parietal junction (rTPJ) ($p < 0.005$, FWE cluster size corrected $p < 0.05$) (Figure 4B, C and Table ST3) displayed the main effect of selection by client: BOLD signal was higher when the participant was not selected (vs. selected) by the client (regardless of the sign of relative merit) when selection by client was revealed as well as during the observation of evidence. The interaction effect was observed in the right posterior superior temporal sulcus (pSTS) at the appraisal stage (Figure S6), anteriorly to the observed rTPJ activity mentioned above, and a distinct part of the temporo-parietal area, classically associated with perception of natural movement and detection of agency[34,35]. The rTPJ tracked the participants' level of influence over the client showing higher activation when the participant was ignored by the client. This preference is in line with previous studies showing that this brain area is more active when others' actions did not match the participant's predictions [36], and when inferring others' intentions from their actions [9]. It may also be the case that this activity is driven by the negative experience of being ignored, as this factor was not mutually exclusive from client appraisal in our design.

When evaluating the interaction model parameters, we observed that negative self-perception, as assessed by the FNE score [14,32], was associated with tendency to follow the defensive strategy of advice giving, i.e. positive value estimated to the Selection by Client parameter in the interaction model. We then hypothesised that FNE scores may also predict the brain activity associated with selection by the client. Fourteen of the participants that took part in the neuroimaging experiment completed the FNE questionnaire. Consistent with our hypothesis, we found that the rTPJ response to selection by the client (ignored vs. chosen, Figure 4B, C) was modulated by participants' individual FNE score (Figure 4D) ($N=14$, $R^2 = 0.52$, $p = 0.003$, positive relation). Sensitivity of rTPJ to Selection by Client increased with participant's inferior self-perception; the difference in rTPJ activity during ignored compared to selected trials was increased with FNE scores.

Previous studies indicated that being more accurate than others (corresponding to positive relative merit) and having being selected over others are strong drivers of brain's reward sensitive network

[13,37,38] . Whole brain analysis revealed that activity in the left ventral striatum (VS) (Table ST4) showed a main effect of relative merit (Figure 5A, top panels) at the appraisal stage, and that the right VS showed a main effect of selection (Figure 5A, bottom panels) at the observation of evidence stage ($p < 0.01$, FWE cluster size corrected $p < 0.05$) (Table ST2). To provide an independent test of the hypothesis that responses to relative merit and selection by client overlapped in anatomical location with previously reported responses to monetary and social rewards and prediction errors [39–41], we employed a region of interest analysis based on a literature based meta-analysis of reward signals in Ventral Striatum[42]. We defined two regions of interest (ROIs) one for the left and one for right VS as spheres around the coordinates indicated by NeuroSynth forward inference maps for the term 'Ventral Striatum' (see Methods, Figure 5B). We examined the beta values for selection by client (ignored/chosen) and relative merit (positive/negative) in the NeuroSynth ROIs, showing that relative merit effect (positive > negative) preceded the selection by client effect (chosen > ignored) (Figure 5C). The effect was observed bilaterally, but with different statistical strengths (see uncorrected maps online: http://neurovault.org/collections/2204/). To further explore the temporal dynamics of VS activity, we examined the time courses from these ROIs. We followed the steps used in previous studies[12] to examine time courses of trials containing multiple jittered stages, aligning each trial stage timings to the stage's mean time. We then performed a GLM in each time point across trials in each participant, regressing the trial-by-trial selection by trial and relative merit sign, and evaluating group effects (Figure 5D). The regression results replicated the temporal order that was obtained by our whole brain GLM approach, by which VS responses tracked relative merit first and selection by client later.

We also found that the client's reward prediction error at the outcome stage were tracked in the participant's left and right VS (Figure S7) [43], i.e. earlier than the observed effect of relative merit. As the reward is associated with the client's reward in the task description, such response could be interpreted as a vicarious reward response [43,44], but may also be associated with tracking one's own advice accuracy following the revelation of the coin's location. It is important to note that the long delay in hemodynamic response makes it hard to lock the timing of the observed effects to a specific stage within the trials. However, the temporal jitters implanted in our design permit inferring the temporal *order* of effects reliably. Striatal activity therefore reflected different social computational components of the interactive scenario as they emerged within one trial, supporting the notion that it has a general role in valuation of various motivational factors that drive social behaviour [38,45].

Finally, we examined the functional connectivity between the brain areas implicated in our previous analysis, which may give rise to the behavioural interaction effect between relative merit and selection by client. We examined the psychophysiological interaction (PPI) [46,47] between mPFC and rTPJ, and between the left VS and rTPJ. We contrasted PPI coefficients contrasting positive and negative relative merit trials. We found a significant connectivity between mPFC and rTPJ (Figure 6), which was higher for positive vs negative merit trials (t(17) = 3.25, p = 0.004). In addition, individual PPI coefficients were significantly correlated with the weight of relative merit on advice deviance (β_{Merit}) estimated by our interaction model (R(17) = 0.66, R^2 = 0.44, p = 0.003). The connectivity between left VS and rTPJ was not significant (t(17) = 0.7, p = 0.5), and the coefficients were not correlated with the behavioural model estimations. Increased connectivity between mPFC and rTPJ during positive merit trials points to a possible mechanism for integrating the information about merit and selection by client which could have given rise to the interaction effect we observed in the behavioural results. Our results show that when merit is positive rTPJ may be more sensitive to being ignored. The fact that this connectivity is correlated with the behaviour as revealed by the estimated model parameters for the merit supports that notion.

Discussion

We set out to study how humans attempt to influence others. We developed a novel laboratory paradigm in which participants attempted to persuade a client to take their advice over that of a competing adviser. We observed a consistent pattern of behavioural result across multiple cohorts of participants that interacted with virtual or real confederates in web-based or lab-based experiments. Advice giving behaviour was driven by the interaction between two factors: the current level of influence over the client and the relative (i.e. social comparative) merit with the rival. Most prominently, when participants' relative merit was positive (i.e. they were doing better than their rival), they followed what we called the 'competitive' strategy [3] by expressing higher confidence when ignored by the client and lower confidence when selected by the client. Inter-individual variability in participants' behaviour was captured by a personalized quantitative combination of the impact of these two factors on advice confidence placing each participant's strategy along a spectrum from competitive[3] to defensive[15]. We found that brain activity in two distinct cortical areas was consistent with tracking the fluctuations of these two factors across the experiment. The mPFC tracked the relative (social comparative) merit and the rTPJ tracked selection by client. The temporal dynamics of the BOLD signal in these areas, and in the VS, were also consistent with the unfolding of events in the course of a trial. Finally, we found that functional connectivity between these two brain areas was modulated by relative merit suggesting a neuronal connectivity hypothesis for strategic advising behaviour.

Our study goes beyond previous investigations of the neurocognitive basis of social influence in a number of key aspects. First, it put the participants in a position to shape and influence others' choices. Most previous studies of social influence invariably concentrated on how we *react* to attempts by others to shape our behaviour (but see [9], where Hampton et al. examined strategic adaptation of behaviour in a two player strategic game). For example Campbell-Mikeljon et al. [28] and Izuma and Adolphs[48] investigated how our preferences change when we are given others' opinion about objects of our desire. Zink et al.[37] and Lignuel et al.[21] investigated how we infer our status in a dominance hierarchy from our competitive and/or cooperative interaction with others. Although manipulation of preferences and establishing of one's social dominance are two very common ways of how humans try to shape each other's behaviour, these example studies did not use them as such. Instead, they focused on how the participants *reacted* to others' attempts to shape their behaviour. For example, Lignuel et al.[21] focused on how, after learning their place in a dominance hierarchy, participants chose their opponents in subsequent encounters to ensure maximum earnings. The alternative scenario, which was not studied, would have involved participants establishing their dominance so that trouble-making opponents would avoid them in future. To our knowledge, only one previous study examined advice giving behaviour [13], but again participants were not able to influence their client's future decisions.

Second, even though many previous works studied human social interactions with more than one confederate, to our knowledge, none placed their participants in a context involving dissimilar (asymmetric) social relationships to the confederates. In many studies social interactions involved tracking another person's reliability or intentions [9,12], or tracking multiple agents all playing a similar role[23–25,28]. Still other studies examined how people evaluated group behaviour, where groups consisted of agents with similar incentive and roles and participants engaged in competitive bidding procedure [49], reaching consensus[50], inferring one's hierarchical rank [21,22] or tracking group's preferences[51]. In our task, as is the case in many real-life scenarios, the participant had to track two distinct types of relationships: a competitive one with a rival adviser, and a hierarchical one with a client whose appraisal they seek. This novel configuration of asymmetric social relationships allowed us to disentangle the separate contributions of the elements of the social brain system [8,10], namely that of rTPJ and mPFC to influencing others. These contributions consisted of an internal inference processes tuned to social comparison and construction of one's relative merit in mPFC [22,23,33] and tracking of externally driven processes tuned to evaluation of social outcomes arising from others' behaviour in relation to us in rTPJ. Finally, one may interpret the selection by client as an exogenous event to which the participant's attention is oriented at the beginning of every trial, in contrast, for example, to the latent process of social comparison tracked in the mPFC. The rTPJ responsivity to exogenous social events would therefore be in line with previous findings about the involvement of rTPJ in the orienting of attention to salient external events [52,53]. This distinction is in line with the previous results examining strategic behaviour in two-person game [9], where mPFC was associated with internal inference processes and the rTPJ with evaluating external signals and other's behaviour.

Both relative merit and selection by client affected ventral striatum BOLD activity, a neural structure implicated in valuation of motivational factors in decision making, in line with their combined interactive impact on advising behaviour. Previous studies have shown that striatal neural activity may encode multiple social attributes such as reputation [39], selection [37], appraisal [13] and vicarious reward [43]. Here striatal activity was correlated with the different computational components of the interactive scenario as they emerged in the course of the interaction: reward outcome, social comparative merit and social appraisal. This finding supports the notion that ventral striatum has a domain-general and dynamic role in valuation of various components of our environment as they unfold [38,45].

Computational analysis of behaviour revealed considerable variations in participants' advising strategies, with each participant's behaviour falling on a continuum from pure 'defensive'[15] to pure 'competitive'[3] strategy. These variations were captured by a personalized quantitative combination of the impact of relative merit and selection by client on advice confidence and were consistent with the social rank theory: participants with negative self-perception, scoring high on fear of negative evaluation, were more likely to follow the defensive strategy[14]. In the neuroimaging experiment, participants with higher scores of negative self-perception displayed increased response in the rTPJ to being ignored by the client. These findings provide converging evidence linking management of social influence to negative self-perception and use of our cognitive framework should help future research on the social basis of mental health disorders such as depression. Deterioration of selfesteem and retreating from social engagement are two of the earliest and most debilitating hallmarks of depression[14]. The laboratory model described here offers a uniquely appropriate ecologically valid tool for measuring these social cognitive characteristics of depression.

Our results demonstrate how people use confidence reports as a persuasive signal in a strategic manner. Such use of confidence reports is in line with the literature of persuasion and information sharing [4,3,54]. However, numerous studies have used similar confidence reports to study the process of metacognition, i.e. the internal process of evaluating one's precepts and decisions [55,56]. Our findings demonstrate how, depending on the social context, confidence reports can depart from simply describing the uncertainty in sensory information or decision variables. The findings underscore the importance of taking extra care about the framing (e.g. experimental instructions) and phrasing of *how* to ask participants to report their confidence in psychological and neuroscientific empirical investigations. Our findings support the view of metacognition as a multi-layered process[57], in which *expressing* one's confidence depends not only on uncertainty at low level sensory processing or decision processes aiming to maximise reward, but is also affected by other systematic, non-trivial sources of variance such as social comparison, closeness and friendship [58,59], and social expectation [60].

People with a more accurate opinion are often more confident. But the converse is not necessarily true: being more confident is not necessarily predictive of accuracy. Many keen observers of human condition (e.g. Bertrand Russell, W. B. Yeats and William Shakespeare, to name but a few) have complained that people who know a lot are fraught with self-doubt while the ignorant are passionately confident. Our behavioural and neurobiological results suggest that passionate overconfidence of the underdog could be better understood as sensible recourse to the competitive strategy designed to gain higher social influence, a behaviour supported by the brain's social and valuation system. The philosophers' sad lamentation therefore highlight the importance of social comparison process shown here, in moderating competitive behaviour of the ignorant.

Acknowledgment

UH and BB are supported by the European Research Council (NeuroCoDec 309865). SP was supported by a Marie Sklodowska-Curie Individual European Fellowship (PIEF-GA-2012 Grant 328822) and is currently supported by an ATIP-Avenir grant (R16069JS). CDF is supported by the Arts and Humanities Research Council (grant AH/M005933/1).

Author Contributions

UH, BB and CDF designed the experiment. UH programmed the task. UH and SB collected the data. UH carried data analysis, model based data analysis was done with SP. UH and BB wrote the manuscript; and all the authors edited the manuscript.

Methods

Participants

We recruited four cohorts of participants for this study. All participants provided an informed consent, and received monetary compensation, both approved by the research ethics committee at UCL. Following a pilot experiment, we estimated our effect size to be around 0.5. As our experiment follows a within-participants design, we decided to recruit 60 participants for the online experiment and 30 participants for the longer lab based experiment. We recruited 60 participants for the online experiment using Amazon M-Turk. Two online participants were excluded from analysis as they did not use the full confidence scale. Online participants included 31 males (ages mean ± std 33.7 ± 9.6) and 27 females (ages 36 ± 8.5). We recruited 30 participants for a lab based experiment, in which participants carried the experiment on computers in the psychology department building. One participant was excluded from analysis as she used only one advice level. Lab participants included 13 males (ages 26.5 ± 6) and 16 females (ages 26.2 ± 6). Finally, 19 participants were recruited for the neuroimaging part of the experiment. These corresponded to the expected effect size of activity in previous social cognition neuroimaging studies, for example [13,37,48]. Of these, one participant was excluded from analysis as his fMRI data was contaminated by head movements. Neuroimaging participants were all males (ages 24.7 ± 6.6). We used only male participants in the fMRI experiment to optimize the homogeneity of the participants, and avoid a possibility of task-irrelevant sex-stereotypical behavior effecting our data [21,61,62]. While our behavioural results did not point to any significant gender effect, following from earlier studies (Lignuel et al. 2016), we opted for this more cautious approach in the neuroimaging data collection. Finally, in the fully interactive experiment (experiment 4) we collected data from 19 triplets (57 participants, 24 males aged 25.2 ± 4.76, and 33 females aged 21.2 ± 2.25), and excluded 3 triplets from analysis. These included 2 triplets in which advisers used only the highest confidence levels (4 and 5), and one triplet in which the client chose only one adviser throughout the experiment. We therefore analysed the data from 32 advisers.

Client and Rival Adviser Algorithms

All participants in the main experiment played the role of an adviser, while the client and the other adviser were played by computer algorithm. The other adviser's advice were calculated on each trial according to the probability of the coin being in the black urn (between 0-1), plus noise ($\sim N\left(0,0.08\right)$), to range between [5W 5B], just like the participants' advice. After outcome is revealed on each trial both advice's prognostic value is calculated by multiplying the confidence level (1-5) by accuracy (indicating the correct coin location, i.e. W for white or B for Black urn, 1 = correct, -1 = incorrect, Eq. 1).

The client's choice of adviser was determined by assigning an influence weight to each adviser, updating the weights after each outcome and choosing the adviser with the higher weight in the next trial. The weights summed to 10 and were set to be 5 for each adviser in the beginning of the experiment. To update the client influence weights, we used prognostic value of advice (*PA*) which were derived from advice according to the following rule: When the black urn is suggested then confidence is between [-5, -1] and *PA* is calculated by confidence +6 to range between [1, 5]. When the white urn is suggested then confidence is between [1, 5] and *PA* is calculated by confidence +5 to range between [6, 10]. We used the notation PA_P for to refer to prognostic value of the participant's advice and PA_O for prognostic value of the other's (rival) advice. The weights were updated after each trial according to the last trials' prognostic values, following a rule similar to the one used by Bayarri and DeGroot [3]:

$$[4] \qquad w_P\left(t+1\right)=10\cdot\frac{w_P\left(t\right)\cdot PA_P\left(t\right)^2}{w_P\left(t\right)\cdot PA_P\left(t\right)^2+w_O\left(t\right)\cdot PA_O\left(t\right)^2}$$

Where w_P is the influence weight assigned to the participant. The other adviser's influence weight was defined as $w_O\left(t+1\right)=10-w_P\left(t+1\right)$. Note that when the influence weight of one adviser increases, the other adviser's influence decreases in the same amount. When both advisers give the same advice the influence weights remain the same.

Procedure

We carried out the main experiment on three different platforms: online, in the lab and in the scanner. The main experimental design features were the same across these experiments with a number of minor differences in implementation.

In the online experiment, participants were recruited using Amazon M-Turk. These participants carried out the experimental task online on their own computers using the mouse to input their confidence rating. They received a fixed monetary compensation and were promised a bonus if the client selected them on more than 100 trials. The online experiment had 130 trials. Evidence stage lasted 500ms, and all other stages of the trial were self-paced. Advice giving stage ended when confidence was reported. After the outcome was displayed, participant proceeded to the next trial by pressing a 'Next' button.

Lab-based participants were invited to the lab in groups of three and were told that they are about to play an adviser game together and that the roles of two advisers and client will be assigned randomly at the beginning of the experiment. The participants were then seated in isolated individual cubicles. Unbeknownst to the participants, all three players were assigned to the adviser role and played against a virtual client and a virtual rival adviser. Lab based participants received a fixed monetary compensation, and did not have any further incentive. Lab-based experiment consisted of four blocks of 70 trials. Advice giving stage was self-paced. All other stages lasted a predefined length of time: Appraisal stage lasted 1.5 seconds, evidence stage lasted 500ms, showdown stage lasted 2 seconds and outcome stage lasted 2 seconds.

In the neuroimaging experiment, participants arrived at the scanner unit and met two confederates and were given the same cover story as the lab-based participants. Participants were then put in the scanner, and were instructed on the use of response boxes for inputting their confidence ratings: left hand response box shifted the rating towards the black urn, right hand response box shifted them towards the white urn. Participants received a fixed monetary compensation. In this experiment intervals between stages were set to 1.5 second plus a jittered interval sampled from a Poisson distribution (range: 0-3 seconds; mean 1.5 second). Neuroimaging experiment consisted of four blocks with 60 trials. Advice giving stage was self-paced. All other stages lasted between 1.5 to 4.5 seconds. In addition, four intervals of 8 seconds rest were randomly dispersed between trials of each run.

In addition to the main experiment, we ran a fully interactive experiment in which all roles in the task, advisers and client, were played by human participants. This experiment was carried in a similar manner to the lab-based main experiment, with participants arriving in groups of three, and carrying the experiment on computers in separate cubicles where they were randomly assigned to the roles of advisers or client. Participants received a fixed monetary compensation and did not get any further incentive. The experiment consisted of 130 trials. The pace of the experiment depended on the interaction between the participants. Advisers waited for the client to choose one of them and then the client waited for both advisers to express their advice.

Selective Manipulation of Advice Quality

In our early pilot sessions, we noticed that sometimes the virtual client did not shift between participants, practically ignoring one of them throughout the experiment. This happened because advisers tended to be very similarly calibrated with the evidence. As we intended to use a within-participants design to compare advice on periods in which the adviser is chosen and periods in which he is ignored by the client, we needed a way to manipulate the probability of switching between selected and ignored status. Therefore, in restricted periods of a block, we introduced some noise to one of the two advisers' evidence such that the ratio of black and white squares in the grid became a poor predictor of the reward location. This procedure went as follows: if on a specific trial the probability of the coin being in the black urn was 0.75, the grid would normally include 75 black and 25 white squares (Figure S1). For example, on a noisy trial, this composition changed to 55 black squares and 45 white squares, akin to a reduction in contrast by 20 squares. In all noisy trials' contrasts were reduced by 20 squares in a similar fashion (Figure S1). The procedure ensured that one advisor's advice accuracy was systematically inferior to the other one for a number of consecutive trials thus increasing the probability that the virtual client would shift to selecting the other adviser.

Model Fitting Procedure

We used models with increased complexity to explain advice deviance reported by the participants. The most elaborated model is the *interaction model*, described in equation [3], which assumes that advice deviance is affected by: systematic bias in confidence, current selection by the client (ignored/chosen), relative merit tracked by comparison with rival (equation 2), and the interaction between relative merit and selection by client. Simpler models included: a *mixture model* that exclude the interaction parameter, a *relative merit model* excluding all selection parameters, a *selection model* excluding all relative merit parameters, and a *bias model* excluding all relative merit and selection parameters. An additional model was also tested that was identical to the 'Interaction' model but used the magnitude and sign of relative merit instead of only the sign of relative merit.

We fitted all models to individual advice deviance. We used a cost function, $L(M)$ to estimate a given model M fit to the data. The cost function compared the advice deviance estimated with the model M ($AdviceDeviance_M(t)$) and the actual advice deviance observed in behaviour on each trial ($AdviceDeviance_{Data}(t)$):

[5] $$L(M) = -\sum_{t=1..T} \log\left(\frac{1}{1+abs\left(AdviceDeviance_M(t) - AdviceDeviance_{Data}(t)\right)}\right)$$

Like log likelihood cost function, the ratio inside the log is close to one when the estimation is close to the data, and it gets closer to zero when the distance between estimation and data increases. Therefore, lower values of the cost function indicate better fit of the model to the data. We used a Markov Chain Monte Carlo (MCMC) Metropolis algorithm for model fitting and estimation for each participant [63–65]. For model comparisons we calculated individual Deviance Information Criterion (DIC) [64], which uses the distribution of likelihood obtained and penalizes for increased number of parameters (Supplementary Figure 2). We used in house Matlab code and the MCMC toolbox for Matlab by Marko Laine (http://helios.fmi.fi/~lainema/mcmc/#sec-4).

Parameter estimation was done individually by using an integrating over the parameter values in the Markov process chain [65]. The learning rate parameters estimated using the Interaction model were used in the aggregated analysis to determine the trial by trial relative merit (equation 1) and separate trials to positive and negative relative merit. The mean parameter estimations for all models are reported in table ST1 in the supplementary materials.

MRI data acquisition

Structural and functional MRI data were acquired using Siemens Avanto 1.5 T scanner equipped with a 32-channel head coil at the Birkbeck-UCL Centre for Neuroimaging. The echoplanar image (EPI) sequence was acquired in an ascending manner, at an oblique angle (≈30°) to the AC–PC line to decrease the impact of susceptibility artefact in the orbitofrontal cortex [66] with the following acquisition parameters: volumes, 44 2mm slices, 1mm slice gap; echo time = 50 ms; repetition time = 3740 ms; flip angle = 90°; field of view = 192 mm; matrix size = 64 × 64. As the time for each block was dependent on the participants' reaction time, overall functional blocks changed in length, and approximately 250 volumes were acquired in about 15 minutes and 40 seconds. A structural image was collected for each participant using MP-RAGE (TR = 2730 ms, TE = 3.57 ms, voxel size = 1 mm3, 176 slices). In addition, a gradient field mapping was acquired for each participant.

fMRI data analysis

Imaging data were analysed using Matlab (R2013b) and Statistical Parametric Mapping software (SPM12; Wellcome Trust Centre for Neuroimaging, London, UK). Images were corrected for field inhomogeneity and corrected for head motion. They were subsequently realigned, coregistered, normalized to the Montreal Neurological Institute template, spatially smoothed (8 mm FWHM Gaussian kernel), and high filtered (128 seconds) following SPM12 standard preprocessing procedures.

We carried two complementary data analyses. Using the computational behaviour analysis, we labelled trials according to selection by client (chosen/ignored) and relative merit (positive/negative) and examined the effect of selection, relative merit and their interaction on brain activity. We used individual level general linear model (GLM) with stick predictors at the onset of each stage (appraisal, evidence, advice report, other advice display, outcome), and additional boxcar predictor 1.5 seconds long ending at the time of advice report confirmation, capturing the motor button presses. We ran GLMs in which the condition labels (Selection by Client (chosen/ignored) and relative merit (positive/negative)) applied to the outcome stage, appraisal stage and evidence stage separately, to overcome the problem of correlation between predictors of interest, as the order of our stages was fixed.

In a separate analysis we examined how the trial-by-trial prognostic values comparison modulated brain activity. We used a set of variables as parametric modulators of activity which included the trial-by-trial prognostic values comparison as the regressor of interest, and aggregated internal status, client's reward prediction error, and unsigned participant's advice as regressors of no interest. We used different GLMs to estimate the effect of this set of parameter modulations on the outcome stage, appraisal stage and evidence stage separately.

In region of interest (ROI) analysis we examined event related effects in specific brain region in different stages within an advice giving trial. Individual beta maps were estimated and sampled within ROIs using MarsBar SPM toolbox. In addition we used NeuroSynth[42] defined ROIs of the left and right ventral striatum. We selected the peak of the ventral striatum reverse inference map, made from 310 studies. We used a 12mm sphere around the left and right peak activity as ROI using MarsBar SPM toolbox (MNI coordinates [-12 8 -8], [10 6 -8], z > 22).

To examine the time course of the changes in brain activity in the regions of interest, we followed previous studies[12] and exploited the time jitters to disentangle the brain activity corresponding to different cognitive processes of interest. We separated each subject's time series sampled from the VS into each trial, and resampled each trial to a duration of 15s: previous trial other's advice (stage 5) at time 0, previous trial outcome (stage 6) at time 2s, Selection by Client (stage 1) at time 4s, evidence (stage 2) at time 6s, confidence report (stages 3-4) at time 9s-11s, and current trial other's advice (stage 5) at time 13s (These timings were the mean timings across all trials in all subjects.) The resampling resolution was 100ms. This temporal realignment allowed the observation of signal throughout the trial while taking advantage of the random jitter and fast event-related design. We then performed a GLM at each time point across trials in each subject. We had one regressor for selection by client, and another one for the sign of relative merit. We then calculated the mean of the effect across subjects at each time point, and their standard errors.

We used psychophysiological interaction (PPI)[67] analysis to examine the functional connectivity between brain areas implicated in the main GLM analysis. The goal of this analysis was to identify patterns of differential connectivity during positive and negative merit trials. The model was estimated in the following steps. We extracted individual average time-series of BOLD activity within the group defined mPFC (**Figure 4**) and left VS (**Figure 5**), which showed Social Comparisons and Relative Merit effects, respectively. The resulting time courses were deconvolved using standard procedures[47]. We estimated a GLM of BOLD responses in rTPJ (**Figure 4**) with the regressors from the standard GLM (see above) and including the PPI regressor: interaction between the mPFC deconvolved time series and an indicator function for outcome stage positive vs negative relative merit, and the mPFC deconvolved time series. These regressors were convolved with a canonical hemodynamic response. A similar GLM was carried with PPI regressor from the left VS at Client Appraisal stage. The individual beta values from for the PPI analysis were used in a second level analysis of group effect, and tested against estimated model parameters.

Questionnaires

As follow up for our experiment we sent participants Fear of Negative Evaluation (FNE) questionnaire [32] to be filled online, to test the link between trait selection perception and social behaviour, as predicted by social rank theory [31]. The questionnaire was sent six months after the main experiment, to make sure there is no effect of the questionnaires on the performance in the task and vice versa. Participants were paid for completing the questionnaires. 62 of our original 106 participants filled the questionnaire, 14 from the fMRI cohort, 15 from the lab cohort and 33 from the online cohort.

Data Availability

The behavioural data that support the findings of this study are available from the corresponding author upon reasonable request. All statistical parametric maps from the neuroimaging part of the experiment are available from NeuroVault: http://neurovault.org/collections/2204/.

References

1. Packard, V. *The hidden persuaders*. (Penguin books, 1991).
2. Cialdini, R. B. *Influence*. **3,** (A. Michel, 1987).
3. Bayarri, M. J. & DeGroot, M. H. Optimal reporting of predictions. *J. Am. Stat. Assoc.* **84,** 214–222 (1989).
4. Sah, S., Moore, D. a. & MacCoun, R. J. Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organ. Behav. Hum. Decis. Process.* **121,** 246–255 (2013).
5. Tetlock, P. *Expert political judgment: How good is it? How can we know?* (Princeton University Press, 2005).
6. Tenney, E. R., MacCoun, R. J., Spellman, B. A. & Hastie, R. Calibration trumps confidence as a basis for witness credibility: Research report. *Psychol. Sci.* **18,** 46–50 (2007).

7. Brosseau-Liard, P. E. & Poulin-Dubois, D. Sensitivity to Confidence Cues Increases during the Second Year of Life. *Infancy* **19,** 461–475 (2014).

8. Koster-Hale, J. & Saxe, R. Theory of Mind: A Neural Prediction Problem. *Neuron* **79,** 836–848 (2013).

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

10. Gallagher, H. L. & Frith, C. D. Functional imaging of 'theory of mind'. *Trends Cogn. Sci.* **7,** 77–83 (2003).

11. Izuma, K. The Social Neuroscience of Reputation. *Neurosci. Res.* **72,** 283–288 (2012).

12. Behrens, T. E. J., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. S. Associative learning of social value. *Nature* **456,** 245–9 (2008).

13. Mobbs, D. *et al.* Reflected glory and failure: the role of the medial prefrontal cortex and ventral striatum in self vs other relevance during advice-giving outcomes. *Soc. Cogn. Affect. Neurosci.* **10,** 1323–1328 (2015).

14. Gilbert, P. The relationship of shame, social anxiety and depression: the role of the evaluation of social rank. *Clin. Psychol. Psychother.* **7,** 174–189 (2000).

15. Price, J., Sloman, L., Gardner, R., Gilbert, P. & Rohde, P. The social competition hypothesis of depression. *Br. J. Psychiatry* **164,** 309–15 (1994).

16. Tetlock, P. E. Accountability: The neglected social context of judgment and choice. *Res. Organ. Behav.* (1985).

17. Greenwald, A. G. The totalitarian ego: Fabrication and revision of personal history. *Am. Psychol.* **35,** 603–618 (1980).

18. Schlenker, B. R. *Impression management: The self-concept, social identity, and interpersonal relations*. (Brooks/Cole Publishing Company Monterey, CA, 1980).

19. Festinger, L. A Theory of Social Comparison Processes. *Hum. Relations* **7,** 117–140 (1954).

20. Wheeler, L. Motivation as a determinant of upward comparison. *J. Exp. Soc. Psychol.* **1,** 27–31 (1966).

21. Ligneul, R., Obeso, I., Ruff, C. & Dreher, J. Dynamical representation of dominance relationships in the human medial prefrontal cortex. *Curr. Biol.* **1,** 1–33 (2016).

22. Kumaran, D., Banino, A., Blundell, C., Hassabis, D. & Dayan, P. Computations Underlying Social Hierarchy Learning: Distinct Neural Mechanisms for Updating and Representing Self-Relevant Information. *Neuron* **92,** 1135–1147 (2016).

23. Wittmann, M. K. *et al.* Self-Other Mergence in the Frontal Cortex during Cooperation and Competition. *Neuron* **91,** 482–493 (2016).

24. Boorman, E. D., O'Doherty, J. P., Adolphs, R. & Rangel, A. The Behavioral and Neural Mechanisms Underlying the Tracking of Expertise. *Neuron* **80,** 1558–1571 (2013).

25. Campbell-Meiklejohn, D., Simonsen, A., Frith, C. D. & Daw, N. D. Independent Neural Computation of Value from The Confidence of Others. (2016). doi:10.1523/JNEUROSCI.4490-15.2016

26. Lebreton, M., Kawa, S., Forgeot d'Arc, B., Daunizeau, J. & Pessiglione, M. Your Goal Is Mine: Unraveling Mimetic Desires in the Human Brain. *J. Neurosci.* **32,** 7146–7157 (2012).

27. Chung, D., Christopoulos, G. I., King-Casas, B., Ball, S. B. & Chiu, P. H. Social signals of safety and risk confer utility and have asymmetric effects on observers' choices. *Nat. Neurosci.* **18,** 912–6 (2015).

28. Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J. & Frith, C. D. How the opinion of others affects our valuation of objects. *Curr. Biol.* **20,** 1165–70 (2010).

29. Smith, D. V., Clithero, J. A., Boltuck, S. E. & Huettel, S. A. Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. *Soc. Cogn. Affect. Neurosci.* **9,** 2017–2025 (2013).

30. Boorman, E. D., Behrens, T. E. & Rushworth, M. F. Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. *PLoS Biol.* **9,** e1001093 (2011).

31. Gilbert, P. The evolution of social attractiveness and its role in shame, humiliation, guilt and therapy. *Br. J. Med. Psychol.* **70,** 113–147 (1997).

32. Carleton, R. N., Collimore, K. C. & Asmundson, G. J. G. Social anxiety and fear of negative evaluation: Construct validity of the BFNE-II. *J. Anxiety Disord.* **21,** 131–141 (2007).

33. Gilbert, S. J. *et al.* Functional Specialization within Rostral Prefrontal Cortex (Area 10): A Meta-analysis. *J. Cogn. Neurosci.* **18,** 932–948 (2006).

34. Mars, R. B. *et al.* Connectivity-based subdivisions of the human right 'temporoparietal junction area': evidence for different areas participating in different cortical networks. *Cereb. Cortex* **22,** 1894–903 (2012).

35. Deen, B., Koldewyn, K., Kanwisher, N. & Saxe, R. Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cereb. Cortex* 1–14 (2015). doi:10.1093/cercor/bhv111

36. Vander Wyk, B. C., Hudac, C. M., Carter, E. J., Sobel, D. M. & Pelphrey, K. A. Action Understanding in the Superior Temporal Sulcus Region. *Psychol. Sci.* **20,** 771–777 (2009).

37. Zink, C. F. *et al.* Know Your Place: Neural Processing of Social Hierarchy in Humans. *Neuron* **58,** 273–283 (2008).

38. Báez-Mendoza, R. & Schultz, W. The role of the striatum in social behavior. *Front. Neurosci.* **7,** 1–14 (2013).

39. Izuma, K., Saito, D. N. & Sadato, N. Processing of social and monetary rewards in the human striatum. *Neuron* **58,** 284–94 (2008).

40. Ly, M., Haynes, M. R., Barter, J. W., Weinberger, D. R. & Zink, C. F. Subjective socioeconomic status predicts human ventral striatal responses to social status information. *Curr. Biol.* **21,** 794–7 (2011).

41. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69,** 1204–15 (2011).

42. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8,** 665–670 (2011).

43. Mobbs, D. *et al.* A key role for similarity in vicarious reward. *Science (80-.).* **324,** 900 (2009).

44. Morelli, S. A., Sacchet, M. D. & Zaki, J. Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. *Neuroimage* **112,** 244–253 (2015).

45. Lebreton, M., Jorge, S., Michel, V., Thirion, B. & Pessiglione, M. An Automatic Valuation

System in the Human Brain: Evidence from Functional Neuroimaging. *Neuron* **64,** 431–439 (2009).

46. Friston, K. . *et al.* Psychophysiological and Modulatory Interactions in Neuroimaging. *Neuroimage* **6,** 218–229 (1997).

47. Gitelman, D. R., Penny, W. D., Ashburner, J. & Friston, K. J. Modeling regional and psychophysiologic interactions in fMRI: The importance of hemodynamic deconvolution. *Neuroimage* **19,** 200–207 (2003).

48. Izuma, K. & Adolphs, R. Social manipulation of preference in the human brain. *Neuron* **78,** 563–73 (2013).

49. van den Bos, W., Talwar, A. & McClure, S. M. Neural Correlates of Reinforcement Learning and Social Preferences in Competitive Bidding. *J. Neurosci.* **33,** 2137–2146 (2013).

50. Suzuki, S., Adachi, R., Dunne, S., Bossaerts, P. & O'Doherty, J. P. Neural Mechanisms Underlying Human Consensus Decision-Making. *Neuron* 1–12 (2015). doi:10.1016/j.neuron.2015.03.019

51. Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A. & Fernández, G. Reinforcement Learning Signal Predicts Social Conformity. *Neuron* **61,** 140–151 (2009).

52. Mitchell, J. P. Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cereb. Cortex* **18,** 262–71 (2008).

53. Corbetta, M., Patel, G. & Shulman, G. L. The reorienting system of the human brain: from environment to theory of mind. *Neuron* **58,** 306–24 (2008).

54. Farrell, J. & Rabin, M. Cheap Talk. *J. Econ. Perspect.* **10,** 103–118 (1996).

55. Bahrami, B. *et al.* What failure in collective decision-making tells us about metacognition. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367,** 1350–65 (2012).

56. Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **329,** 1541–3 (2010).

57. Shea, N. *et al.* Supra-personal cognitive control and metacognition. *Trends Cogn. Sci.* **18,** 186–93 (2014).

58. Almaatouq, A., Radaelli, L., Pentland, A. & Shmueli, E. Are you your friends' friend? Poor perception of friendship ties limits the ability to promote behavioral change. *PLoS One* **11,** 1–13 (2016).

59. Pontari, B. A. & Schlenker, B. R. Helping Friends Manage Impressions: We Like Helpful Liars But Respect Nonhelpful Truth Tellers. *Basic Appl. Soc. Psych.* **28,** 177–183 (2006).

60. Bowles, H. R., Babcock, L. & Lai, L. Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask. *Organ. Behav. Hum. Decis. Process.* **103,** 84–103 (2007).

61. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol.* **27,** 415–444 (2001).

62. Buchan, N. R., Croson, R. T. a & Solnick, S. Trust and gender: An examination of behavior and beliefs in the Investment Game. *J. Econ. Behav. Organ.* **68,** 466–476 (2008).

63. Haario, H. An adaptive Metropolis algorithm. *Bernoulli* **7,** 223–242 (2001).

64. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **64,** 583–639 (2002).

65. Kruschke, J. *Doing Bayesian data analysis: A tutorial introduction with R JAGS, and Stan*. *Igarss 2014* (Elsevier, 2015).

66. Weiskopf, N., Hutton, C., Josephs, O. & Deichmann, R. Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. *Neuroimage* **33,** 493–504 (2006).

67. Friston, K. J. *et al.* Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6,** 218–29 (1997).

Figures

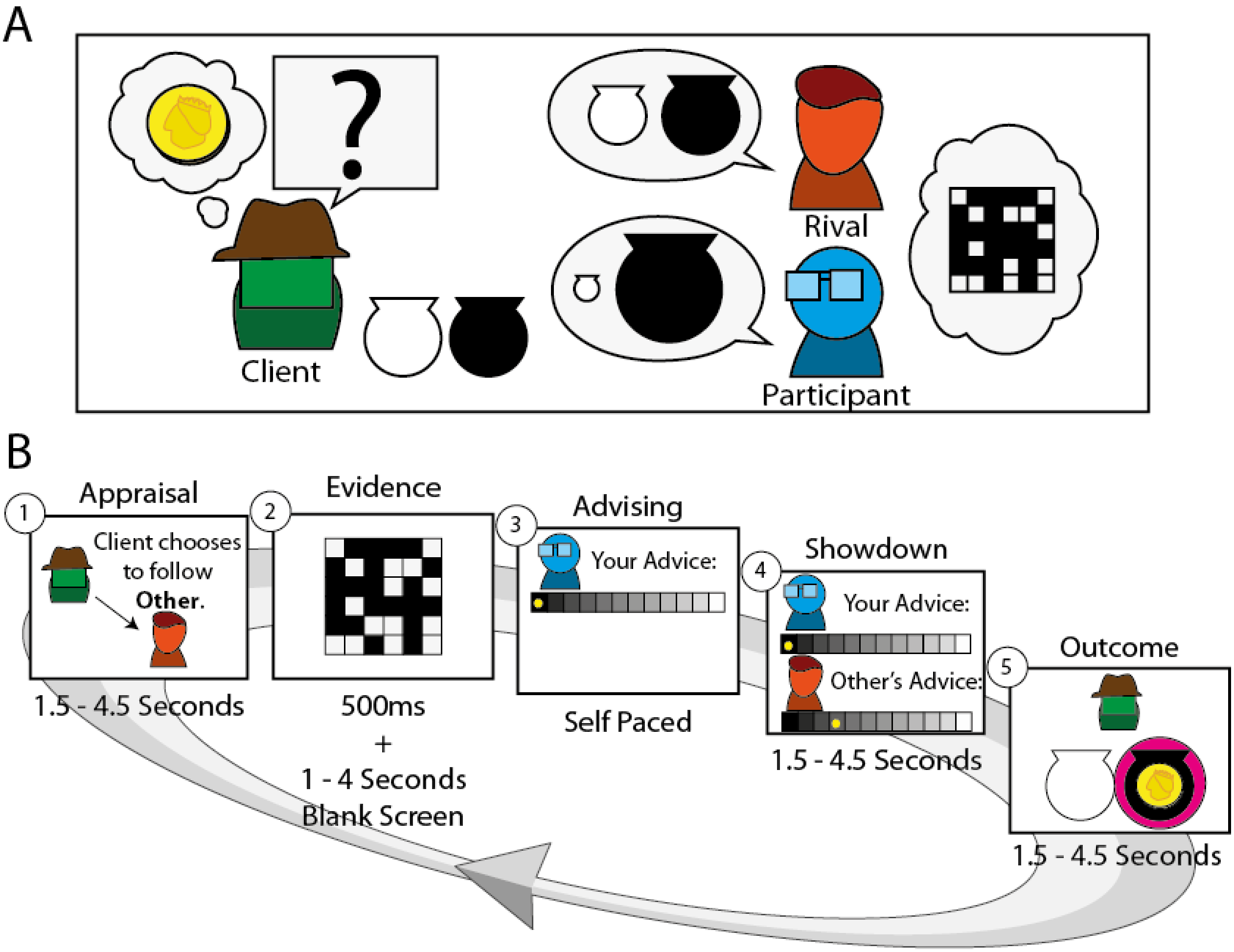

Fig. 1: Experimental Design

(A) Participants were engaged in an advice giving task. In this task a client is looking for a coin hidden in a black or white urn. He relies on advice from two advisers (the participant (blue) and a computer generated adviser (red)). The advisers, but not the client, have access to information regarding the probability of the coin location. The client considers the advisers' previous success and current confidence when choosing an adviser to follow on each trial. (B) Each trial contains five stages. (1) Appraisal: In the beginning of each trial the client chooses the adviser he is going to follow on the commencing trial (and consequently which adviser is ignored). (2) Evidence: The participant (and the rival) then sees a grid of black and white squares, whose ratio represents the probability of the coin being in the black urn. (3) Advising: The participant states his advice on coin location using a 10 levels confidence scale ranging from “definitely in the black urn” (5B) to “definitely in the white urn” (5W). (4) Showdown: both advisers’ opinion is displayed. (5) Outcome: The content of the urn suggested by the selected adviser (magenta circle) is revealed. The next trial starts with appraisal by client based on the history of confidence and success. The stages timings indicated are from the fMRI experiment, where jitter was introduced between stages. See supplementary materials for online and lab stages timings. A playable demo of the experiment can be viewed at http://www.urihertz.net/AdviserExperiment.html .

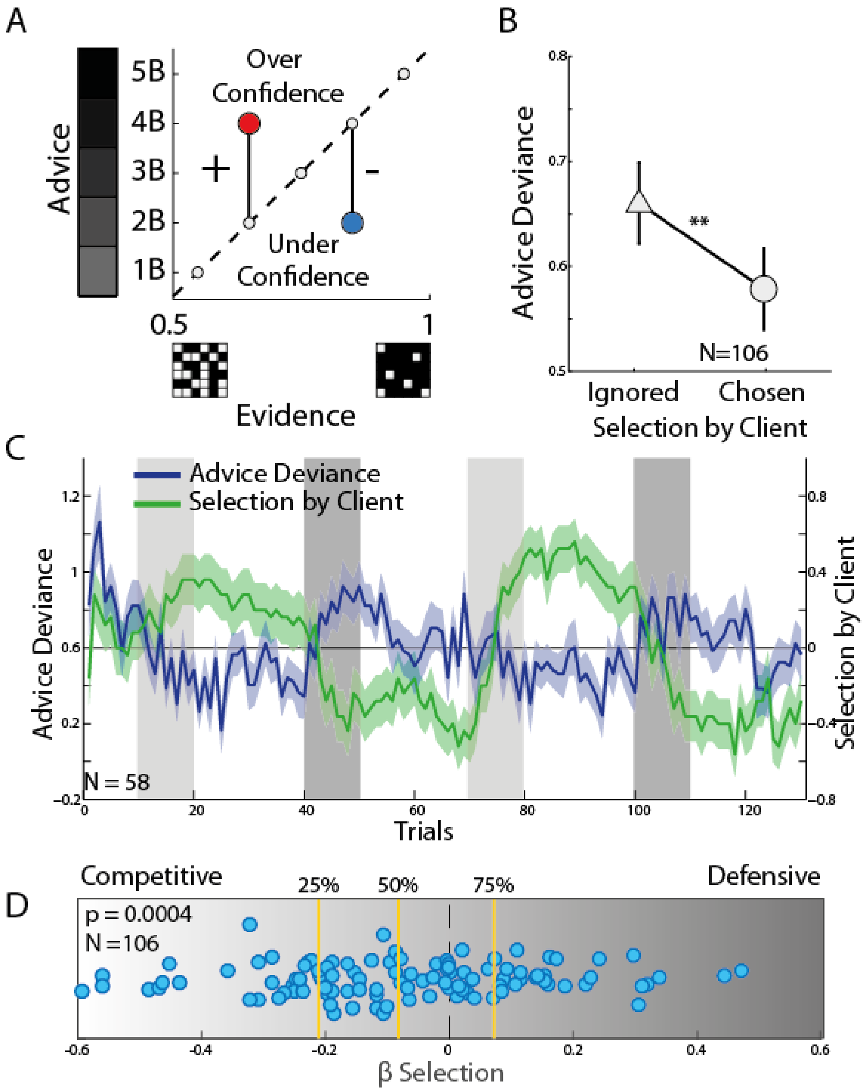

Fig. 2: Client's Selection Effect on Advice

(A) Deviance of advice from probabilistic evidence. Confidence is plotted against evidence grid, i.e. probability that coin is in the black urn. The dashed line represents zero-deviance policy in which confidence matches the ratio of black to white squares in the evidence grid. Overconfident advice (red circle) would lie above the dashed line and was defined as positive deviance; conversely, underconfident advice (blue circle) corresponded to negative deviance. (B) Averaged advice deviance under high vs low selection (paired t-test, t(105) = 3.45, p = 0.001). (C) The dynamic interplay of selection and advice deviance in the online experiment. Blue line = trial-by-trial average advice deviance. Green line = average selection. Light green and blue = standard error of the means. Light and dark grey blocks mark periods of low evidence reliability for the rival and for the participant, respectively (see Fig. S1). (D) Our best fitting model contained a parameter governing how Selection by Client affect advice deviance, ‘β-Selection’. Positive values of β-Selection are associated with the defensive strategy, i.e. increased advice deviance when being selected. Negative values are associated with the competitive strategy, i.e. decreased advice deviance when being chosen by the client. Despite high level of individual differences, this parameter was significantly lower than 0 across participants (p < 0.0004), in line with our previous analyses.

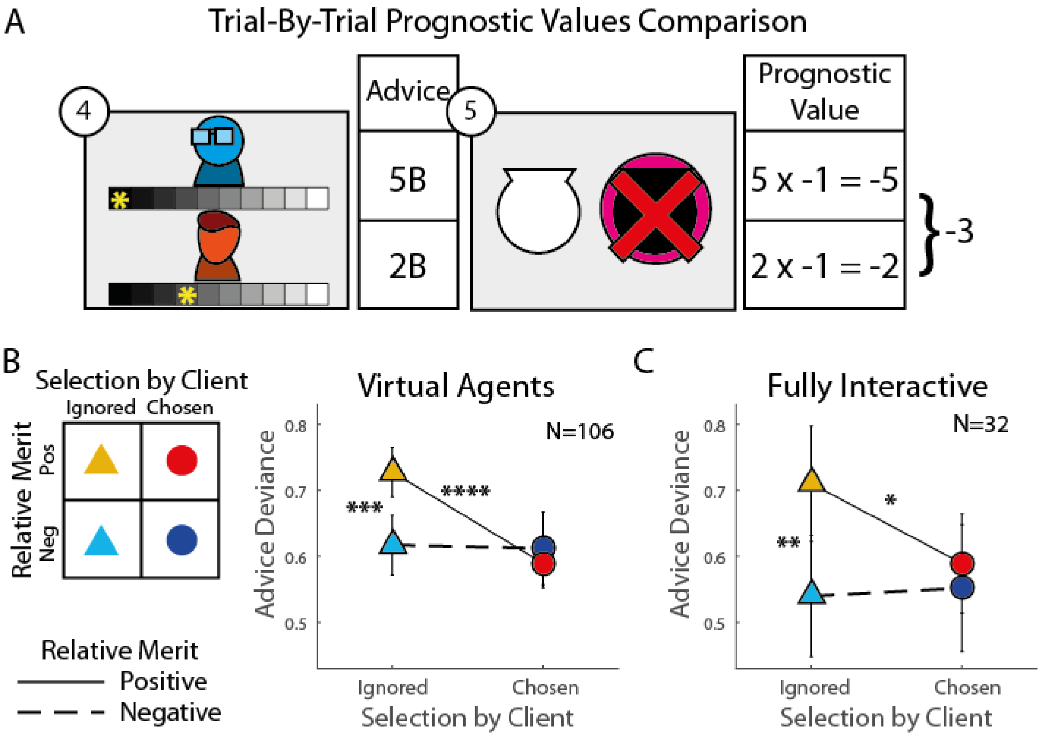

Fig. 3 Relative Merit Effect on Advice

(A) At the end of each trial participants could evaluate and compare their and the rival’s advice prognostic value. The difference in prognostic value was accumulated to form the participants’ relative merit. (B) Our best fitting model included a free parameter for accumulation rate for differences in prognostic values, for relative merit and for the interaction between merit and client’s selection. We used the individual estimated parameters to evaluate the trial by trial relative merit, and divided the trials to four conditions according the Selection by Client and relative merit. Advice deviance was highest (i.e. most overconfident) when participants’ positive relative merit conflicted with the client's choice to ignore them, and demonstrated a significant interaction effect. (C) The same pattern of behaviour was replicated in the fully interactive experiment, where three human participants played the roles of a client and two advisers. (two-tailed paired t-test comparisons: * $p < 0.05$, ** $p < 0.005$, *** $p < 0.0005$, **** $p < 0.00005$, Error bars indicate SEM).

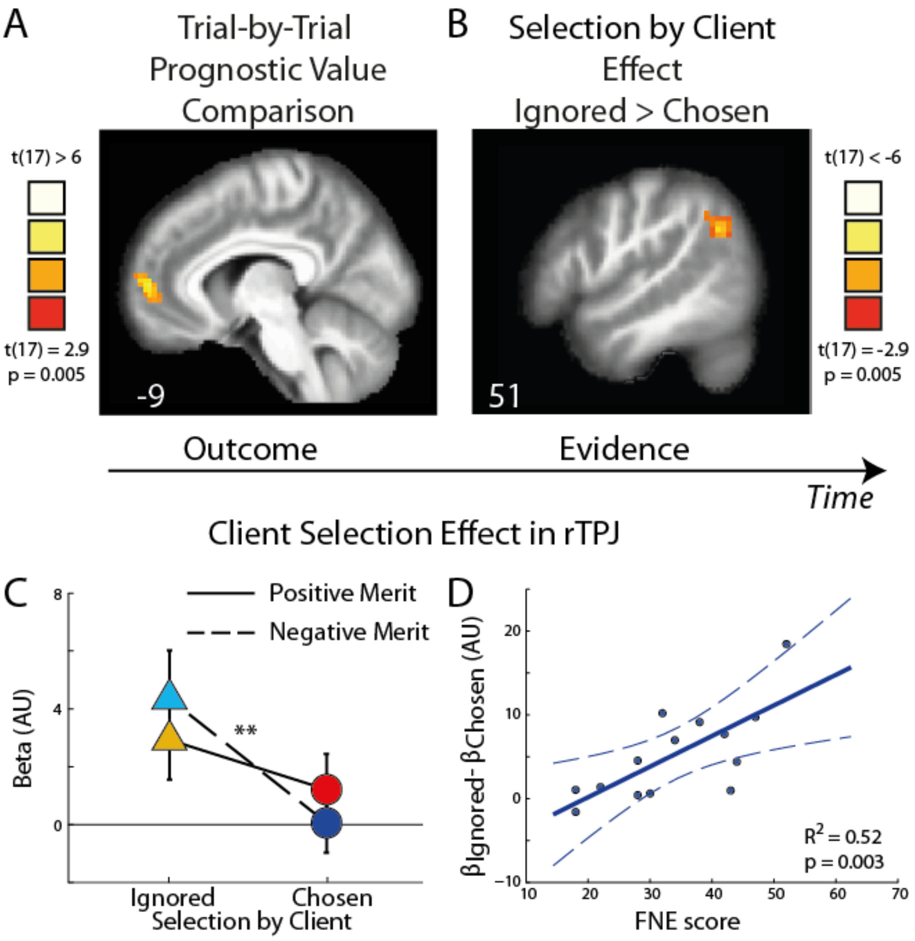

Fig. 4: Encoding of Relative merit and Selection in the cortex

(A) During the outcome stage, before appraisal value was revealed, activity in mPFC (MNI coordinates [x,y,z]: [-9,56,5]) was positively modulated by the trial-by-trial prognostic value comparison (Figure 2D, Eq. 1) ($p < 0.005$, FWE cluster size corrected $p < 0.05$). (B) At the observation of evidence stage, activity in right TPJ (coordinates: [42,-58,35]) was higher on ignored vs chosen trials ($p < 0.005$, FWE cluster size corrected $p < 0.05$). (C) Analysis of beta estimates from the right TPJ showed a significant Selection by Client effect (N = 18, F(1,53) = 16.26, p = 0.0009). Error bars indicate SEM. (D) Client selection effect on activity in the right TPJ was correlated with the participants' fear of negative evaluation (FNE) score: differences in rTPJ activity between ignored and chosen trials was greater for participants with high FNE score (N=14, as we obtained FNE scores from a subset of the participants). Regression line is computed from the population-level estimate of the FNE scores on selection by client effect in the rTPJ. Dashed lines indicate 2 SEs.

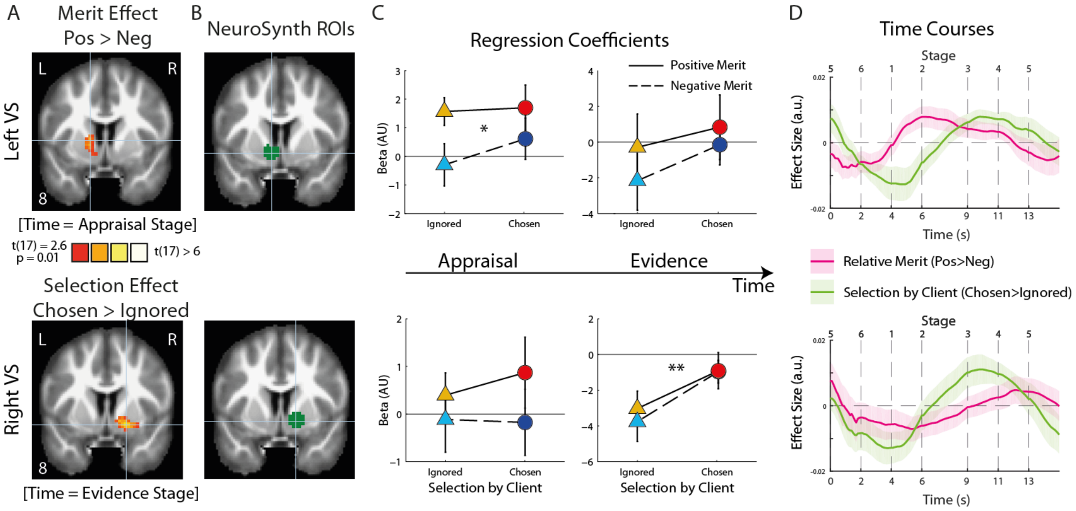

Fig. 5: Encoding of Relative merit and Selection in the Striatum

(A) Whole brain analysis showed that, at the appraisal stage, activity in left VS (coordinates: [-15,8,5]) was higher in positive relative merit trials compared to negative relative merit trials (top panel, p < 0.01, FWE cluster size corrected p < 0.05). At the evidence stage activity in right VS (coordinates: [18,8,-10]) was higher in high selection trials compared to low selection trials (bottom panel, p < 0.01, FWE cluster size corrected p < 0.05). (B) ROIs of left and right VS defined by Neurosynth meta-analysis. (C) Analysis of beta estimates sampled from the Neurosynth defined left and right VS. Left VS showed a significant relative merit effect at appraisal stage (F(1,53) = 6.96, p = 0.017, top left panel). The right VS showed a significant selection effect at the evidence stage (F(1,53) = 14.42, p = 0.0014, bottom right panel). Error bars indicate SEM. (D) Time course of the effects of Relative Merit (pink) and Selection by Client (green) are shown across all stages of a trial, from the showdown stage (5) of a preceding trial to the outcome stage (6) of the current trial. These time courses demonstrate the temporal order of coding of Relative Merit and Selection by Client in VS. Thick lines: mean effect size. Shadows: SEM.

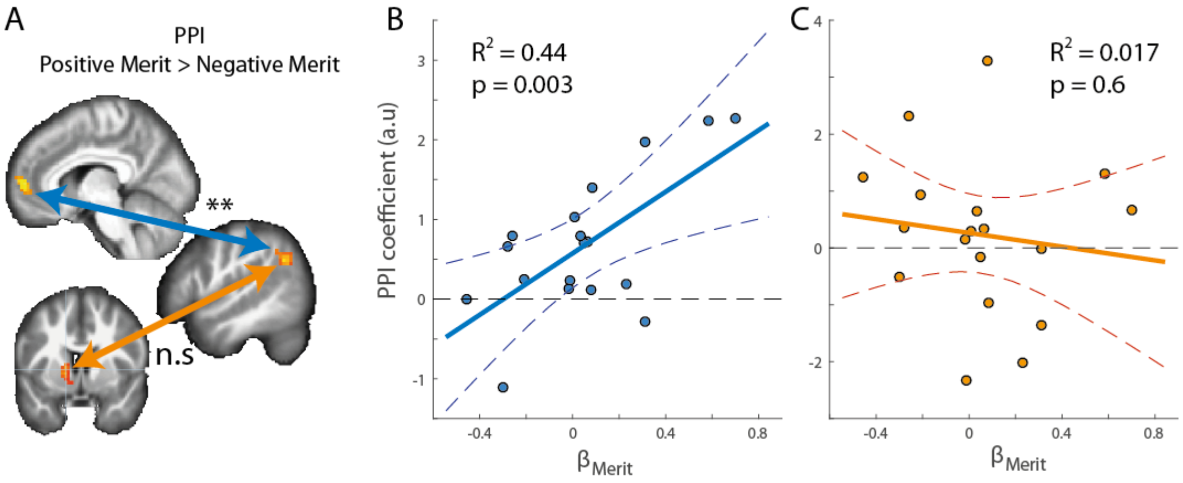

Figure 6: Relative Merit Dependent Functional Connectivity between mPFC and rTPJ

(A) To examine the neural substrates of the interaction between Relative Merit and Selection by Client, we used psychophysiological interaction (PPI) method to assess the connectivity between the mPFC and rTPJ, as well as between left VS and rTPJ. We found a significant merit dependent functional connectivity between mPFC and rTPJ (blue arrow, $p < 0.005$). Functional connectivity between these areas increased during positive (compared to negative) merit trials. Functional connectivity between left VS and rTPJ was not significant. (B) Functional connectivity strength (PPI coefficient) between mPFC and rTPJ was highly correlated with the behavioural effect of merit on advice deviance, captured by our model (β_{Merit}) (N=18). (C) Functional connectivity strength (PPI coefficient) between the left VS and rTPJ was not correlated with the behavioural effect of merit on advice deviance. Regression line is computed from the population-level estimate of β_{Merit} effect on PPI coefficient. Dashed lines indicate 2 SEs.

Supporting Information for:

Neural computations underpinning the strategic management of influence in advice giving

Uri Hertz, Stefano Palminteri, Silvia Brunetti, Cecilie Olesen, Chris D Frith, Bahador Bahrami

Supplementary Figures

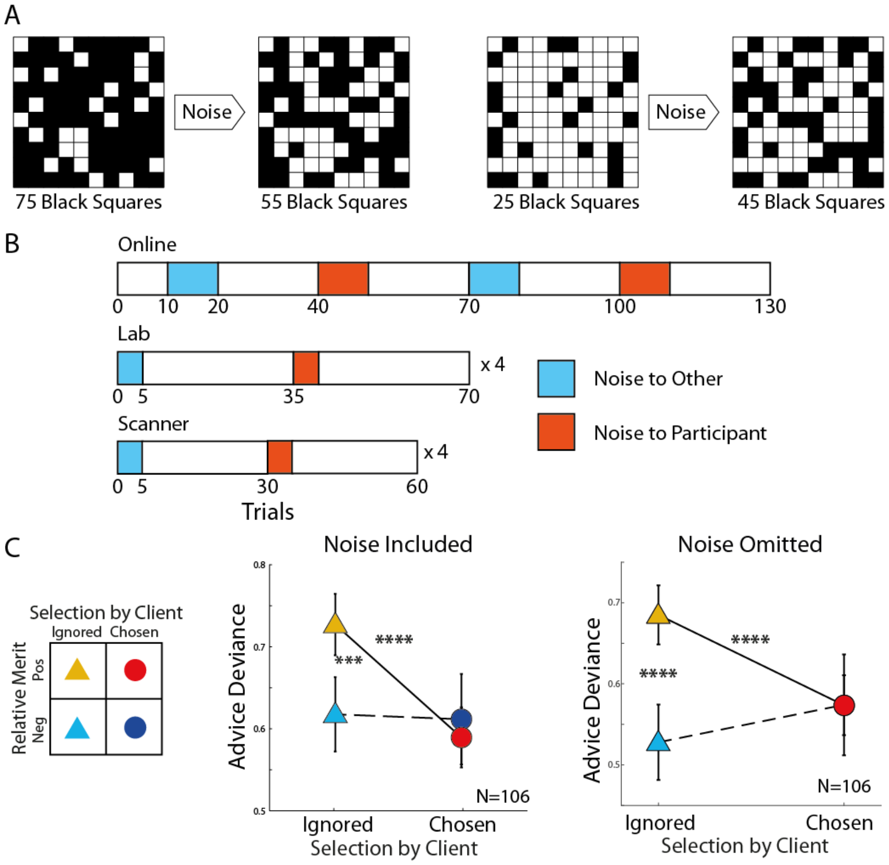

Figure S1: Selective Manipulation of Advice Quality

We used a manipulation to increase the probability of differences in advice and accuracy between advisers, which would then entail client switching between advisers. (A) We therefore introduced noise to one of the advisers' evidence, i.e. the ratio between black and white squares in the grid. The noise procedure went as follows. If the probability of the coin being in the black urn on a specific trial was 0.75, the grid would normally include 75 black squares and 25 white squares (right side of panel A). On a noisy trial, this composition changed to 55 black squares and 45 white squares, i.e. reduction of contrast by 20 squares. Similarly, when the probability of the coin being in the *white* urn on a specific trial was 0.75 (0.25 probability of being in the black), the noisy grid would include 55 white squares (45 black squares) instead of 75 white squares (25 black squares) in the not noisy case, again reducing the contrast by 20 squares (left side of panel A). In all noisy trials' contrasts were reduced by 20 squares in a similar fashion. (B) Noisy periods were relatively short, lasting 10 consecutive trials in the online experiment and 5 consecutive trials in the lab and scanner experiments. Noise was introduced either to the participant or to the other adviser (i.e. the virtual rival algorithm was fed noisy evidence). Online experiments included 4 blocks of noisy periods, 2 for each adviser, while the longer lab based and scanner experiments included 8 blocks of noisy period, 4 for each adviser. (C) We analysed the data before and after omitting the noise periods. Our results did not change after omitting the noisy trials (compare two panels), with a significant main effect of Relative Merit (F(1,315) = 5.29, p = 0.02), and significant Relative Merit x Selection by Client effect (F(1,315) = 13.1,p = 0.0005). Error bars indicate SEM.

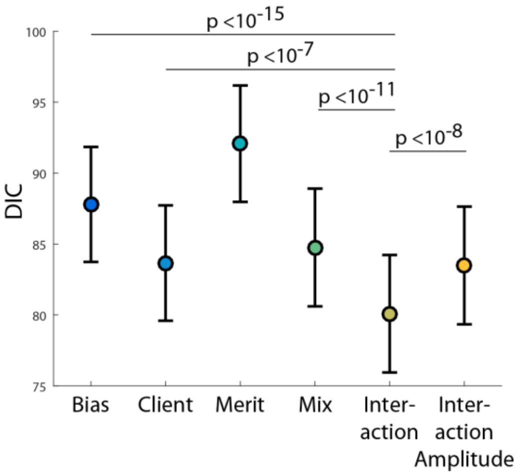

Figure S2: Models Comparison

Mean and SEM of the models' DIC scores, as well as p values for two-sided paired t-test comparisons. We fitted our models using a Markov chain Monte Carlo (MCMC) Metropolis algorithm. This process resulted in likelihood distribution across parameter space that minimizes the individual log likelihood (likelihood of advice deviance given the model's parameter estimation). We calculated individual Deviance Information Criterion (DIC) [36], which uses the distribution of likelihood obtained and penalizes for increased number of parameters. All our models included a Bias parameter to capture trait overconfidence or under-confidence bias. Our baseline model included no other parameters (Bias model), while other models included a Selection by Client parameter (Client), a Relative Merit Parameter (Merit), both Selection by Client and a Relative Merit parameters (Mix), and an additional Interaction parameter (Interaction). An additional model was tested which was identical to the 'Interaction' model but used the magnitude and sign of relative merit instead of only the sign of relative merit (see Methods). Best fit to participants' behaviour was obtained using the Interaction model (see main text). See all estimated models parameters in table S1.

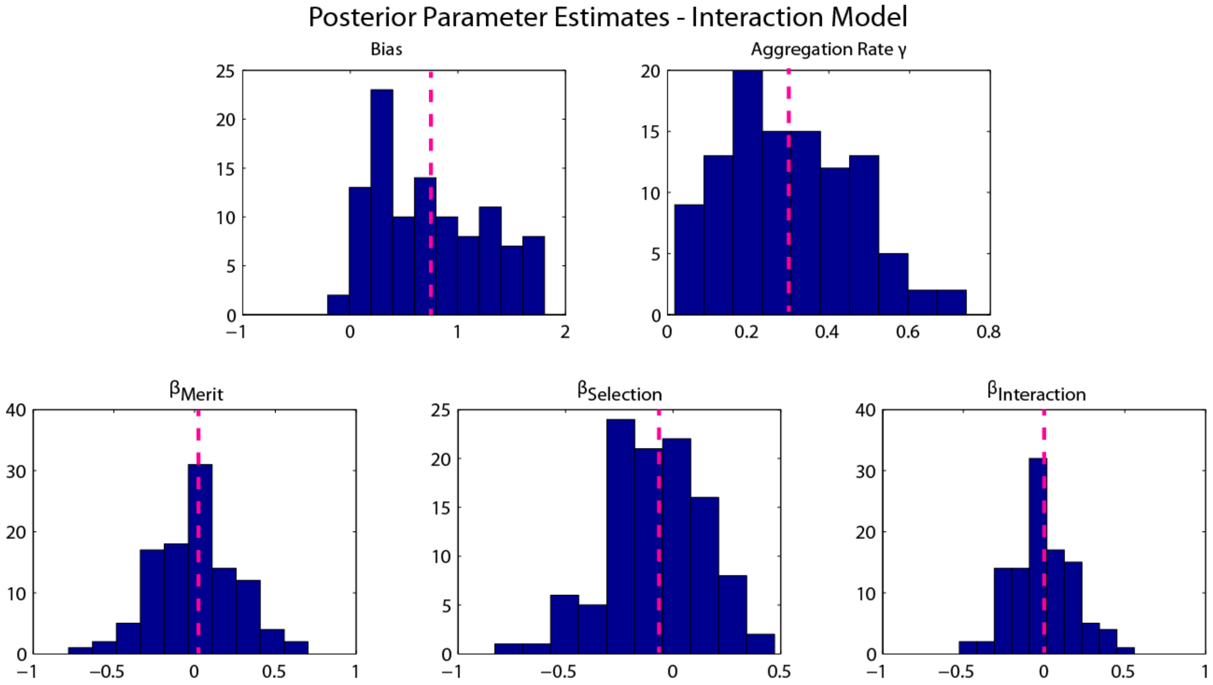

Figure S3: Distributions of Individual Estimated Parameters from the Interaction Model

Our model fitting procedure provided distribution of parameter estimation for each participant. We calculated the posterior individual parameters of the Interaction model to examine individual differences in weights assigned to client selection, relative merit and their interaction. In all figures the histogram of posterior parameters across participants is displayed (n=106), with dashed line marking the mean parameter value. All our participants were overconfident (trait > 0), and had below 0.8 aggregation rate γ. Weight assigned to relative merit was distributed around zero and not significantly different from 0. The weight assigned to client selection was variable but significantly lower than zero (mean = -0.08, p = 0.0004). Weight assigned to the interaction of relative merit and selection was distributed around zero and not significantly different from 0.

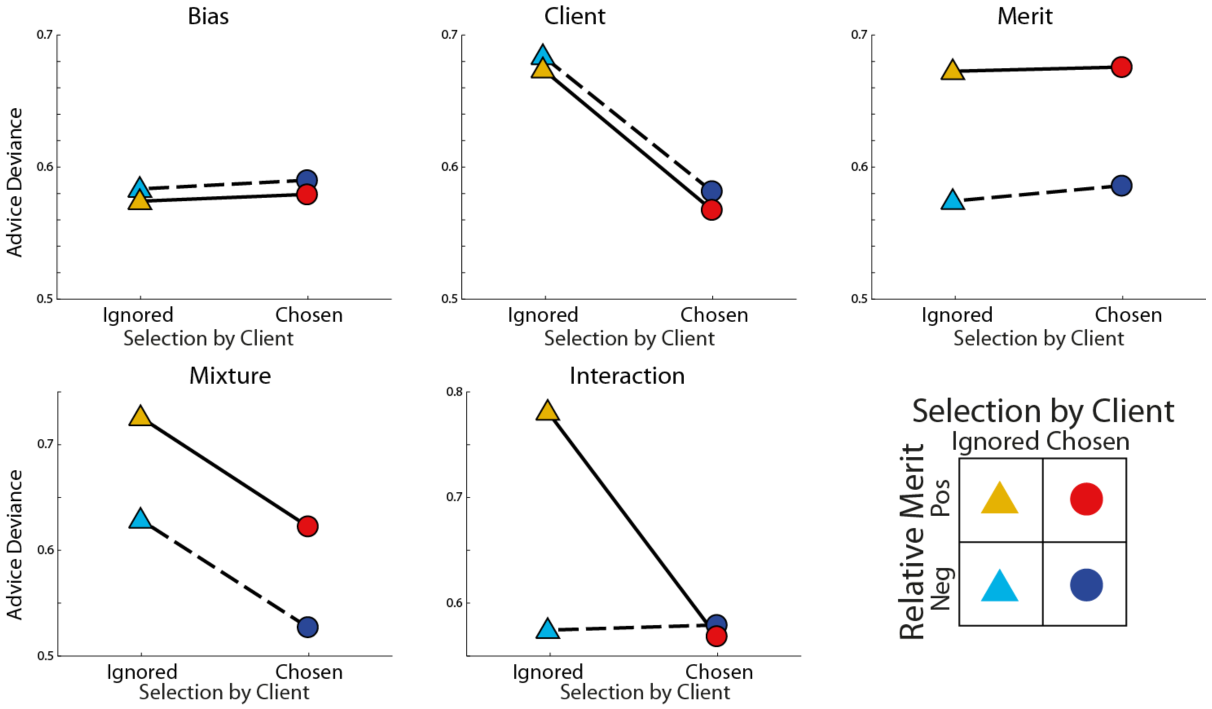

Figure S4: Models Simulations

We ran simulations of our five models and examined how they qualitatively differ from one another, and how well they capture patterns in the data. We used fixed values for free parameters, and the other adviser advices and coin location probabilities from the real obtained data, and estimated advice deviance declared by the participants according to the different models. These simulations show that only a model that take into account client selection, relative merit and their interaction can reproduce the pattern of results observed in participants' advice deviance (Figure 2E).

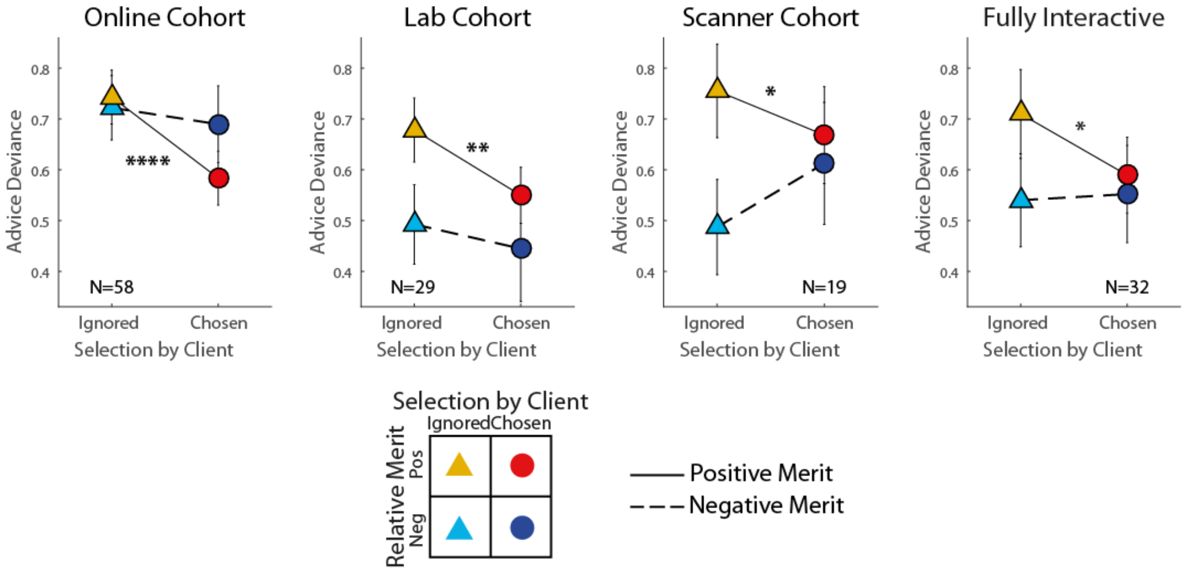

Figure S5: Client Selection and Relative Merit Effect in the Four Cohorts

Aggregated advice deviance was analysed independently in our three cohorts of participants. The main result holds in all cohorts: advisers gave more determined advices when the client chose the rival and ignored them and their relative merit was positive, compared to when the client chose them and they had positive relative merit (two-tailed paired t-test comparisons, * $p < 0.05$, ** $p < 0.005$, **** $p < 0.00005$). Error bars indicate SEM. However, deviance in negative relative merit condition varied across cohorts. We carried a mixed effects ANOVA (2x2x4) to examine the cohort dependent variations with selection, relative merit and Cohort as main fixed effects, and participants as random effect nested in Cohort. This resulted in significant main effect for relative merit (F(1, 412) = 15.5, p = 0.0002), but not main selection effect (F(1, 412) = 3.52, p = 0.06) or Cohort effect (F(2, 412) = 1.12, p = 0.34). As expected from the repeated main result, interaction between relative merit and selection was significant (F(1, 412) = 15.75, p = 0.0001). Interaction between relative merit and Cohort was significant (F(2, 412) = 5.4, p = 0.001), but not interaction between selection and Cohort (F(2, 412) = 0.88, p = 0.45) and not the triple interaction between relative merit, selection and

Cohort (F(2, 412) = 0.21, p = 0.88)). The main result of relative merit and selection interaction sustained across cohorts and experimental settings, including the fully interactive experiment.

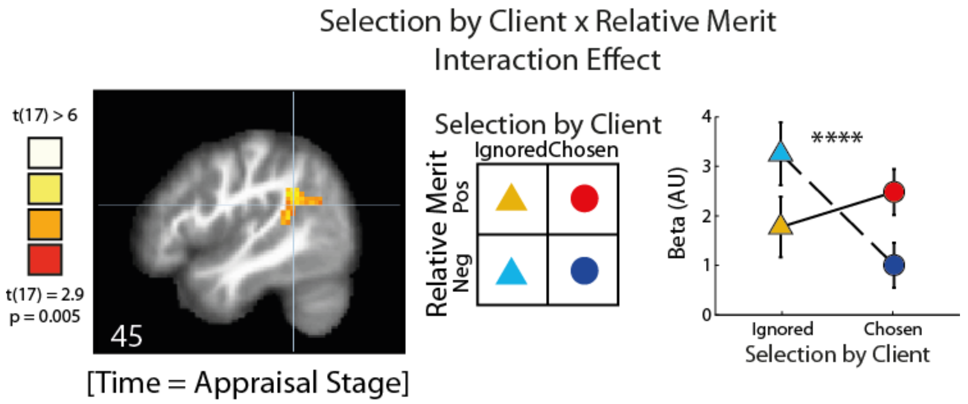

Figure S6: Client Selection and Relative Merit Interaction effect in the superior temporal sulcus (STS)

Whole brain analysis showed that activity in right posterior STS was higher when relative merit and client selection matched compared to mismatch trials during the appraisal stage ($p < 0.005$, FWE cluster size corrected $p < 0.05$). Analysis of beta estimates from the right posterior STS showed a significant client selection x relative merit interaction effect (F(1,53) = 30.24, $p < 0.0001$). Error bars indicate SEM.

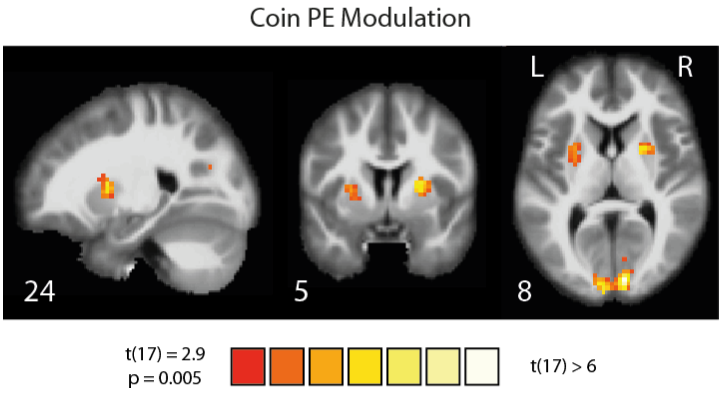

Figure S7: Encoding Client's reward PE during Outcome Stage

Client's reward prediction error was calculated as the difference between the probability of the coin being in the black urn and its actual location, multiply by reward (-1 for no coin and 1 for coin).

[4] $$ClientRewardPE(t) = \left|CoinInBlack(t) - p(Black)\right| * Reward(t)$$

During the outcome stage activity in right and left Striatum was positively modulated by the coin PE (the sign of surprise of coin location) ($p < 0.005$ Uncorrected).

Supplementary Tables

Model	Parameters					DIC
	$Bias$	γ	$\beta_{Selection}$	β_{Merit}	$\beta_{Interaction}$	
Bias	0.68±0.06					87.79±4.05
Client	0.72±0.06		-0.07±0.02			83.65±4.06
Merit	0.96±0.10	0.48±0.015		-0.36±0.07		92.07±4.1
Mix	0.74± 0.06	0.57±0.01	-0.08±0.03	0.006±0.02		84.75±4.15
Interaction	0.74±0.05	0.3±0.015	-0.08±0.02	0.003±0.02	-0.008±0.02	80.08±4.14
Interaction Amplitude	0.72±0.05	0.24±0.001	-0.08±0.02	0.02±0.02	0.009±0.01	83.48±4.15

Table ST1 - Model's parameters and fitting scores (Mean ± SEM)

				MNI Coordinates		
Sign	Region Name	Extent	t-value	x	y	z
Positive	Superior Medial Gyrus	87	4.8	-9	56	5

Table ST2 – Modulation of activity during outcome stage by trial-by-trial prognostic value comparison threshold at p < 0.005 (Relates to Figure 4) which survived cluster size FEW correction (p < 0.05). See full maps in NeuroVault: http://neurovault.org/collections/2204/

				MNI Coordinates		
Sign	Region Name	Extent	t-value	x	y	z
Negative	R Angular Gyrus	88	4.31	42	-58	35
Positive	R Putamen	85	-4.46	18	8	-10

Table ST3 – Activations in 'Chosen' > 'Ignored' client selection contrast during evidence stage, threshold at p < 0.005(Relates to Figure 4,5) which survived cluster size FEW correction (p < 0.05). See full maps in NeuroVault: http://neurovault.org/collections/2204/

				MNI Coordinates		
Sign	Region Name	Extent	t-value	x	y	z
Positive	L Putamen	82	3.62	-15	8	5

	L Superior Orbital Gyrus	82	3.39	-15	44	-13

Table ST4 – Activations in 'Positive Relative merit' > 'Negative Relative merit' contrast during appraisal stage threshold at p < 0.01 (Relates to Figure 5) which survived cluster size FEW correction (p < 0.05). See full maps in NeuroVault: http://neurovault.org/collections/2204/