Chapter Title: Frames of reference in human social decision making.

*(to appear in **Handbook of Motivational and Cognitive Control**, MIT Press, eds. Mars, Sallet, Rushworth and Yeung)*

Authors: Laurence T. Hunt[1,2] and Timothy E.J. Behrens[1,3]

Author Affiliations:

1. FMRIB Centre , University of Oxford, John Radcliffe Hospital , Headington, Oxford , OX3 9DU
2. Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OX1 3UD
3. Wellcome Trust Centre for Neuroimaging, 12 Queen Square, London, WC1N 3BG

Author e-mails: lhunt@fmrib.ox.ac.uk**;** behrens@fmrib.ox.ac.uk

Word count: ~5,000

The ability to place the study of the neural basis of motivational and cognitive control in its full social context could be considered a fairly recent one. The number and complexity of human social interactions is unique[15], and so it has only been with the advent of tools for neuroimaging that we can begin to understand the physiological correlates of human social behaviour in the brain. The field of social cognitive neuroscience has grown rapidly, and many studies have begun to investigate human subjects' brain activity whilst they choose how to behave in a given social context.

In this chapter, we argue that understanding the *frame of reference* in which neural activity changes in these studies is fundamental to understanding the function of that activity. This point has long been appreciated in brain regions that are involved in sensorimotor integration (e.g. [10; 35]); the challenge here being to elucidate the intermediate transformations between neural activity in one frame of reference (sensation) being transformed into activity in an entirely different one (motor control). We contend that it is equally important to try to establish the frame of reference in which neural activity varies during a social interaction. Similar transformations, from the observation of a social partner's behaviour, to the inference of intention, to the eventual modification of one's own actions, must be performed by the brain to elicit successful social behaviour. Although the discrimination made here may be rather more blunt, two broad categories of activations can be delineated from fMRI studies of social interaction conducted thus far.

A particular social setting will often cause adaption of behaviour in a particular way. This influence of others on one's own behaviour affects activity in a network of brain regions associated with reward and selecting one's *own* actions. However, social interactions also require inference of another individual's intentions, so as to adapt one's own behaviour accordingly. This process of mental inference, or *mentalizing*, engages a quite distinct set of brain regions. It is useful to think of this activity as being in the frame of reference of the *other* person's actions. Successful social behaviour requires *both* inference of other individuals' thoughts *and* adaptation of one's own behaviour accordingly, and several recent studies have investigated activity during tasks in which both processes are required, but in which the two vary *independently* of one another. By ensuring that task variables in the two frames of reference remain independent, variance in neural activity can be ascribed to one, other, or both frames of reference.

Socially-derived reward

It has long been argued that social exchanges can be described in terms of costs and benefits to the individual[25] and so might be valued against other (non-social) costs and rewards. Socially rewarding or aversive stimuli might therefore need to be translated into the same neural 'common currency' as other (non-social) stimuli, within which currency their value can be compared. Evidence for this exists at the level of single neurons; macaques will forego a juice reward to view

a photograph of certain conspecifics[13] and the subjective value of this viewing is reflected in the firing rate of neurons in the lateral intraparietal area at the time of choice[29].

In human subjects, primary rewards and decision values are frequently found to activate portions of ventral striatum and ventromedial prefrontal cortex (see chapter XX (O'Doherty) for review). Recent evidence has been found for common activations in these regions for reward derived from a social interaction.

Ventral striatum activates for reward derived from a social interaction

Izuma and colleagues[26] used functional MRI to scan human volunteers as they received appraisals from social partners on their moral character, and (in a separate condition) as they performed a monetary gambling task. Both monetary rewards and positive social appraisals were found to activate an overlapping portion of the striatum. It may also be rewarding to cooperate with other individuals or to conform to the opinion of experts; the striatum is again active in both cases[39; 9].

A rewarding stimulus, by definition, is one that increases (or reinforces) behaviour that elicits the stimulus. For example, the potential value of social conformity can reinforce behaviours that comply with a social norm. Subjects who adjust their behaviour to comply with others have greater activation in the striatum than those who leave their behaviour unchanged[9]. This striatal activation occurs at the time when the discrepancy between one's own behaviour and other individuals' behaviour is revealed, even if norm compliance is measured at a

much later timepoint in the experiment[30]. Activations in striatum can therefore be related not only to reward, but also to signals observed in paradigms of reinforcement learning, where an error in prediction of reward frequently witnessed in the striatum (e.g. [36]), and is thought to cause an adaptation of one's future behaviour (see chapter XX for review). The concept of a socially-derived reward prediction error has been examined more explicitly in the case of a trust game, in which the striatum is more active on trials where donations from a partner induce an increase in trusting behaviour[27]. Similar to the reward prediction error in dopamine cells[42], striatal signals in this study were found to transfer back in time from the timepoint of the rewarding stimulus (trusting behaviour from the partner) to the timepoint of trial onset.

Ventral striatal activations for other-regarding preferences

Human actions in a social environment are not only guided by self-interest, but may also be influenced by the impact of these actions on the behaviour of others. Altruistic behaviour such as charitable donation has provided a puzzle for pure 'self-interest' economic theories that view donation as giving away a good for no direct gain for oneself in terms of material wealth or reputation. However, such behaviour can be explained by economic models that include terms that incorporate a value for 'other-regarding' preferences[19]. These values show a common neural substrate to that of self-regarding preferences, as indicated by the value of a charitable donation being reflected by BOLD fMRI responses in ventral striatum and ventromedial prefrontal cortex[33; 23; 24].

A direct measure of altruistic tendencies can be obtained by studying the behaviour of individuals participating in collaborative (or competitive) games. Such games may be designed such that certain actions can only be interpreted in terms of their impact on other players, rather than on oneself. By making the games single-shot and anonymous, the potential confounds of later reciprocation and the desire to build a reputation with the social partner are also removed, and so the behaviour of subjects (and their brain activity) can only be explained by appealing to subjects' 'other-regarding' preferences. In one such game[12], subjects could punish unfair behaviour in social partners who retained an unfair share of money with which they had been entrusted by the subject. The striatum was found to be more active when this punishment was costly to the partner rather than symbolic, and was particularly active in subjects who invested a greater amount of their own money in punishing their partner. This set of activations may appeal to the idea that humans punish unfair behaviour *because* they find it rewarding to do so – in other words, the subjects place a higher value on delivering a punishment than in keeping monetary reward to themselves. A similar conclusion can be drawn from the finding that male subjects have a greater striatal activation when they observe an unfair player receiving an electric shock than a fair player receiving the same treatment[44].

Behaviour in these games can be captured by several different quantitative economic models, including those featuring a term that factors in aversion for inequality between participants[18], or those that assume that subjects act in response to fair (or unfair) behaviour with reciprocation[17]. Neural evidence

for the former has come from a recent study[48], that split subjects into unequal starting positions ('rich' or 'poor'). Activation in striatum was stronger for reward delivered to the *partner* in subjects who were in an initial position of being richer than their partner, but stronger for reward delivered to the *subject* in subjects who were in an initial position of being poorer. The findings might be taken as neural evidence that subjects find a reduction in inequality rewarding.

The pitfalls (or promises?) of reverse inference

Does the increased activity in striatum truly reflect the rewarding properties of other-regarding preferences? Striatal activations can also be found for noxious stimuli[3] or monetary losses[43]. This highlights the danger of making a 'reverse inference' about the psychological process being recruited by a given task[37]. We should not leap to the conclusion that because a region is activated in a task (e.g. a behaviour in a social setting), and that region is also known to respond to a particular psychological state (e.g. reward), that this implies that the task must engage that state (i.e. the pro-social behaviour *must be* rewarding). This conclusion would only be valid if the region *solely* responded to reward; in the case of the ventral striatum, this is not the case.

On the other hand, there are cases where reverse inference *has*, apparently, been of use, as in a recent study of why subjects overbid in auctions[14]. When subjects participated in a socially competitive auction, ventral striatum showed a different pattern of activity as compared to a formally identical (but non-socially competitive) lottery. *Winning* in the auction elicited the same

striatal response as winning in the lottery, but *losing* elicited deactivation selectively in the competitive auction. The authors interpreted this finding as subjects showing a particular *aversion to losing to others* in the competitive environment of an auction. Crucially, they then used this reverse inference to design a further behavioural experiment; subjects were found to further increase their overbidding if the auction was framed in terms of losing a certain amount that had already been given to the subject, as opposed to winning the same quantity at the end of the auction.

So, a reverse inference may perhaps be a valid tool if it used to generate a novel hypothesis that can then be tested behaviourally. It is important to bear in mind that without subsequent testing of the generated hypotheses, it is dangerous to leap to conclusions about why a region is found to be more active in a given task.

Socially-derived reward and other-regarding preferences have been studied in the 'self' frame of reference

Whether or not we can make inferences about reward, it is clear that social preferences activate a common set of regions as those engaged in processing more basic rewards. Clearly, this activity is an important aspect of how social influences modify behaviour. Importantly, however, all of the reviewed findings can be considered as reflecting one's *own* utility function for a particular outcome or behaviour. Activity in striatum (and other regions that respond to reward or decision value) may represent a 'final output' of social influences on one's own

behaviour, but this need not imply that the striatum is involved in the *computation* of these social influences in the first place. The key test would be to devise a way to manipulate the factors that *lead* to a particular social influence being expressed, without manipulating the *value* of that social influence to one's own actions.

Action selection and cognitive control in a social environment

It is often the case that other-regarding and self-regarding preferences will come into conflict. A network of brain regions encompassing anterior cingulate cortex (ACC) and dorsolateral prefrontal cortex (DLPFC) has been isolated as showing increased activity when task demands are heightened or competing action plans must be selected between[16; 38] (see chapter XX for review). This same network of regions has also been found to be active in tasks where subjects must weigh the benefits to themselves with the impact their actions will have on another individual.

DLPFC, ACC and 'self-control' in social interactions

Competing motives for self and other have been captured in several economic games (see [7] for a review (or chapter XX Coricelli?)). One such game is the 'ultimatum game', in which two subjects undergo a one-shot interaction in which money provided by the experimenter must be divided. One subject, the 'proposer', makes an offer of how the pot should be split between the two

players, and the second player, the 'responder', decides either to accept this offer (in which case the players receive this split) or reject it (in which case both players leave with nothing). The rational strategy (assuming both players act in their own self-interest) is for the proposer to make the minimum possible offer (an 'unfair' split of the pot), and for the responder to accept this. In practice, modal behaviour is for proposers to offer a fair (50%-50%) split of the pot, and unfair offers (e.g. 20%-80%) are frequently punished by the responder with rejection[47]. In rejecting an unfair offer, the responder sacrifices his own earnings to ensure that the proposer walks away with nothing, which again can be explained by appealing to an aversion to inequality[18] or a desire for reciprocity[17] (but see [51] for evidence of an additional, non-other regarding explanation of this behaviour).

Sanfey and colleagues[41] measured BOLD fMRI in subjects playing the role of responder in several one-shot trials of the ultimatum game, with interleaved trials played with either a human or computer partner. Unfair offers from human partners were typically equally likely to be accepted or rejected, whereas fair offers (or any offer from the computer, fair or unfair) were invariably accepted. DLPFC, ACC and insular cortex were all found to be more active when responders were faced with an unfair proposal than a fair one, and specifically when receiving this unfair proposal from a human. Activity in these areas was therefore greater on trials where subjects were equally likely to accept or reject the offer (that is, when self-regarding and other-regarding preferences came into conflict) than when they were certain to accept the offer (when no such conflict existed).

The importance of DLPFC when balancing prosocial and selfish behaviour has been further emphasised by studies in which its activity has been temporarily disrupted using repetitive transcranial magnetic stimulation (rTMS). After receiving 15 minutes of rTMS to the right (but not left) DLPFC, subjects show a reduced tendency to reject unfair offers in the ultimatum game[32]. The right DLPFC, it is argued, appears critical in overriding the 'selfish' desire to keep an unfair offer when weighing this against an other-regarding rejection. This has been further investigated in a related game, the 'trust' game, in which subjects choose whether to return or keep money with which their partners have entrusted them. rTMS to right DLPFC does not affect subjects' willingness to return the money in a condition where they are doing so anonymously (and where return rates are relatively low). However, whereas most subjects increase their rate of return when their returns are made public, subjects who have received TMS keep the same rate of return as in the anonymous condition[31]. The key difference in the latter condition is that subjects have an opportunity to build a reputation for being trustworthy; subjects who have undergone TMS give the appearance of neglecting this opportunity and only taking into account their own short-term self-interest in making their choices.

'Self' and 'other' frames of reference in ultimatum and trust games

What aspect of behaviour has changed in these subjects? It might be that they are no longer able to infer the impact that their actions will have on the behaviour of other individuals, or what the prevailing norm for fair behaviour is.

Alternatively, it might be that these functions are perfectly intact, and that the *implementation* of this knowledge in modifying their *own* behaviour has changed. In fact, in both studies there was clear evidence for the latter. In the ultimatum game rDLPFC TMS subjects would still report an offer a 80%-20% split as being highly unfair, in line with reports of control subjects - but would nevertheless go on to accept it[32]. Similarly, in the trust game the subjects' reports of what constituted fair behaviour remained in line with those who had not received rTMS to rDLPFC[31]. Impairing the function of rDLPFC does not appear to change one's perception of other individuals, or the impact of one's own behaviour of other individuals; it instead changes the use of this knowledge in guiding one's own behaviour.

This leaves open the possibility that another brain region, or set of regions, supports the inference of one's own behaviour on others. A hint towards which regions might be important in this process was provided by a recent manipulation of the trust game by van den Bos and colleagues[49]. In this game, an *investor* chooses whether to entrust the subject being scanned (the *trustee*) with some money. This money is multiplied by the experimenter; the trustee must then decide how to split it between the two players. In one manipulation the actions of the investor were particularly beneficial to the trustee, but not to himself. This created more conflict for the trustee (in his *own* frame of reference), who is inclined to be fair as he has done better out of the situation. The trustee's rDLPFC is found to be more active. In another manipulation, the actions of the investor were particularly beneficial to *himself*. This changes the inferred

intention of the investor in sending the money, and so is in the frame of reference of *his* actions. This manipulation affected BOLD fMRI responses at another region, which has been emphasised by a complementary literature in social interactions – the right temporoparietal junction.

In the next section, we introduce this region as one of several whose activity may vary as a function of the inferred behaviour of other individuals in a social interaction.

Inferring the intentions of others in social interactions

Much work has gone into studying regions of the brain that support the ability to infer the intentions of other individuals, or to possess a *theory of mind*. One hypothesis[6] argues that there might be a specialised set of brain regions devoted to social cognitive functions including theory of mind. The research effort in this field has been intense, in part spurred on by the finding that autistic patients have a specific deficit in understanding other's intentions, as demonstrated by the 'false belief' task[1]. A dorsomedial portion of prefrontal cortex and portions of the superior temporal sulcus and temporoparietal junction are both more active in tasks requiring inference of false beliefs, (e.g. [20; 21; 50]), and these regions show altered activity in autistic subjects[2].

What metric is coded in these regions that causes them to be important in intentional inference? Whilst it has been established that these regions are typically more active when performing an interactive game with a human partner

as opposed to a computer[40], the precise computations performed by these regions are such a paradigm are only beginning to be elucidated[4].

Dorsomedial PFC activity reflects both the depth of strategic inference and learning about this quantity

Whilst in the trust and ultimatum games decisions can be based purely upon reaction to the partner's behaviour, other games have been devised that require careful consideration of how the partner is *likely* to behave in order to devise one's own strategy. Such thinking can become 'hierarchical' or 'higher-order' – as my strategy will depend upon what my partner is thinking, but this will depends upon what he thinks I am thinking, and so on. Two recent studies have measured neural activity in games requiring this strategic inference.

An elegant means of measuring strategic inference is to make use of a game from experimental economics known as the 'beauty contest'[34]. In this game, a group of subjects have to pick a number between 0 and 100, but the winner is selected by picking the subject who is closest to a fraction *M* (e.g. 4/5) of all other players' selections. If a subject assumes everyone else is naïve and chooses 50 on average, then it makes sense to select 4/5 of 50, 40 ('first-order' theory of mind). A more sophisticated subject might realise that this is what everyone else will think, and so select 32 ('second order' theory of mind). Yet more sophistication will yield an answer of 26, 21, 17, and so on.

Across multiple rounds of the game with varying values of *M*, subjects tend to show a consistent level of strategic inference, but there is considerable

inter-individual variability in the strategy chosen. Coricelli and colleagues[11] exploited this to investigate which brain structures showed differential brain activity across subjects with different levels of strategic sophistication. When contrasting trials in which the game was played with human (as opposed to computer) partners, subjects with higher level of reasoning showed more activity in the same dorsomedial portion of prefrontal cortex that had previously been isolated in the studies of theory of mind.

The level of strategic inference has typically been found to be quite limited in the beauty contest game (the median level being either first- or second-order inference[8]). However, this may be because such calculations must be performed explicitly as opposed to implicitly in this task, and critically, because there is no opportunity to adapt one's strategy in light of witnessing social partners' recent behaviour. Yoshida and colleagues have recently developed a paradigm that allows for precisely this. The task is based upon the 'stag hunt' game[45], in which players must choose whether to act in a cooperative manner with their partner. Cooperative behaviour from both players is more rewarding than non-cooperative play, but unreciprocated cooperative play is least rewarding of all. One therefore needs to infer that the partner is likely to cooperate before deciding that cooperation is worthwhile.

[FIGURE 1 ABOUT HERE PLEASE]

In the version of the stag hunt presented by Yoshida et al., subjects played iteratively with a computer player who adopted a strategy with a particular level of inference. The authors had previously constructed a model[52] which

attempts to infer this strategy from the behaviour of the partner. Human subjects were found to be successful in tracking the level of inference of quite sophisticated computer agents (such as a computer playing a fifth-order strategy). Moreover, when witnessing the partner's behaviour, the dorsomedial prefrontal cortex was found to correlate with a model parameter that described the entropy of the distribution over possible values that the partner's strategy might take[53]. By finding a correlate of this term, which is critical for updating the likely future behaviour of the partner, in DMPFC, we can conclude that DMPFC may play an important role in *learning* about sophisticated aspects of the future behaviour of other individuals.

Regions implicated in theory of mind are implicated in learning via reinforcement about other agents' behaviour

Two further studies have used the strategy of applying a reinforcement learning RL model to tracking the behaviour of a partner in a socially interactive setting. In one study[5], subjects had to simultaneously learn about which of two options was likely to be rewarded, but had the advice of a confederate at each trial, who had the option of providing the subject with the correct (or incorrect) answer. The confederate was motivated such that he might provide helpful or unhelpful advice, but the subject could only learn this motive by carefully observing how often the confederate was helpful. This learning could be tracked using an RL model, which contained separable terms for the *prediction error* and *learning rate* of the confederate's intentions. At the timepoint critical for learning, the DMPFC,

right TPJ and superior temporal sulcus were found to correlate with the prediction error term, which was not in the traditional frame of reference of reward to oneself, but instead in the frame of reference of the *other individual's actions*. A gyral portion of anterior cingulate cortex correlated with the learning rate in this frame of reference.

[FIGURE 2 ABOUT HERE PLEASE]

Another study employed an iterated inspection game, in which an 'inspector' chooses whether or not to monitor the behaviour of a 'worker'[22]. Inspecting is costly if the worker is already working, whilst working is costly if the inspector fails to inspect. If both players were to play the task optimally, the best strategy would be to adopt a mixed strategy of assigning a certain probability to each action, and selecting from these probabilities at random. However, if either player is suboptimal, human subjects might track the *previous* behaviour of the partner, and use this to *infer* a strategy that exploits the other subject's behaviour. A yet more sophisticated strategy would incorporate the *influence* of each player's current action on the next move that the partner would take. Quantitative RL models can be built that deploy each of these strategies; both superior temporal sulcus and DMPFC signal the 'influence update term' at the time critical for learning, and activity in DMPFC correlates with the likelihood that the sophisticated influence model is being used.

The frame of reference has been established, but the computations performed remain unclear

Whilst the use of reinforcement learning models in tracking the value of one's own possible behaviours has now become commonplace (reviewed in chapter XX), these studies argue that the neural mechanisms supporting the tracking of other's behaviour may require similar computations, but implemented in parallel in discrete neural structures, and discrete frames of reference. Whilst they all agree that the computation performed in DMPFC is critical for inference about the intention of a social partner, already some differences can be seen between the terms in the model used to describe activity in this region. For instance, DMPFC activity in the study by Behrens et al.[5] reflected a *signed* prediction on the probability of the social partner *lying*, whereas activity in the study by Hampton[22] reflected an *unsigned* prediction error, that is activity was greatest when the partner's behaviour was most *surprising*. In the Yoshida et al.[53] study, the *entropy* of the distribution over possible partner strategies should determine how important each new partner move is for updating future estimates of partner behaviour, and this term is reflected in the DMPFC; in study by Behrens et al.[5], however, the analogous metric that is important for learning (the *volatility* of subject behaviour) was reflected in the gyral portion of ACC, not DMPFC. Careful dissection and examination of the differences between the tasks and the models that are used to explain these data and to capture the dynamics of the task are needed. This may be achieved by designing tasks in which computational models of activity in these regions can be directly pitted against one another[46].

What about striatal activations for social prediction errors?

Here it is worth briefly revisiting studies that have found signals that resemble prediction errors in a socially interactive setting, but have found these in striatum rather than the theory of mind network. King-Casas and colleagues[27; 28] have carefully examined data collected from subjects interacting in an iterated version of the trust game, that allows for the building of a reputation between investor and trustee. In the trustee's brain, they find increased activity in the head of the caudate nucleus when the investor reciprocates their past behaviour in a generous fashion ('benevolent' reciprocity) as compared to trials when they fail to do so ('malevolent' reciprocity). This activity could be a prediction error in the frame of reference of the *investor's* future behaviour – an adjustment of the trustee's expectations of the investor – or alternatively could be a prediction error in the frame of reference of the *trustee's* future behaviour – as benevolent reciprocity is more likely to induce an increase in trust. King-Casas *et al*. show clear evidence for the latter proposition – the activity in striatum is increased selectively on trials in which the trustee is to increase his *own* level of trusting behaviour in future rounds.

Klucharev and colleagues[30] scanned subjects as they rated the attractiveness of photographs of individuals in a 'hot or not'-style task; they then presented the average rating of a group of other individuals who had rated the picture. As expected, later ratings of the same photographs were highly influenced by what others thought of the photo, and the striatum and ACC were both found to be influenced by conflict between one's own opinions and that of others. Again, however, this signal (likened by the authors to a prediction error) is

in the frame of reference of one's own behaviour, as evidenced by the fact that it is stronger when one's own behaviour is modified by the conflict than when it is not.

Conclusions

There is a rapidly expanding literature on decision-making, or the use of motivational and cognitive control, in a social setting. Socially-derived reward and conflict appear to activate similar networks to those found in non-social settings. The formal, quantitative models derived from game theory allow close measurement of the impact that other's presence or behaviour has on one's own actions. Recent studies have begun to tease apart networks where the social preferences expressed in these games might be computed. Regions of the theory of mind network perform computations in the frame of reference of a social partner's intentions. The right DLPFC appears particularly important in balancing conflict between the inferred intentions of other individuals and one's own goals. By being precise and formal about the metrics that vary within a task and their frame of reference, we should achieve a much more rigorous and precise understanding of the underlying computations that these regions perform.

Questions for future research

- How can the frames of reference in which neural activity varies be refined to accurately reflect neural activity? How do computations in subregions of the

theory of mind network differ from each? Do they vary in discrete frames of reference?

- How is information in discrete frames of reference combined to support action selection? Can we distinguish competing computational accounts of activity within the same region?
- What happens at the single cell, rather than the metabolic, level? What are the computations instantiated at the neural network level that underlie the observed phenomena in the BOLD fMRI signal? Are these computations uniquely human, or do they also take place in other model organisms?

Further reading

Behrens, T. E., Hunt, L. T. & Rushworth, M. F. The computation of social behavior. Science 324, 1160-4 (2009)

Fehr, E. & Camerer, C. F. Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci (Regul Ed) 11, 419-27 (2007)

Yoshida, W., Seymour, B., Friston, K. J. & Dolan, R. J. Neural mechanisms of belief inference during cooperative games. J Neurosci 30, 10744-51 (2010)

Figure Captions

Figure 1. Different levels of strategic inference yield different strategies in experimental games. (a) In the beauty contest game (ref. 11), subjects select a number between 1 and 100, aiming for a fixed proportion of the selections of other players of the game (in this case, 4/5). Assuming other players are naïve (a

first-order strategy) yields an optimal guess of 40; assuming other players are first-order yields an optimal 'second-order' strategy (32), and so on. (b) In the stag hunt game, subjects either hunt a stationary rabbit (small squares) or cooperate with a partner (circle) to hunt a moving stag (large square). The colour intensity of each of the squares shows the *value function* for each position on the board, derived from the computational model devised by Yoshida and colleagues (ref. 52). Adopting a first, second or third-order strategy yields different value functions, and so different movements from the centre of the board (arrows). Fitting the model to real subjects' behaviour allows inference of which strategy best describes subjects' current play.

Figure 2. Different prediction errors in dissociable frames of reference. The figure depicts an example trial from the study by Behrens and colleagues (ref. 5). (a) The subject aims to maximise reward obtained by choosing which cup (left or right) contains money. From previous trials he has learnt a prediction that the right cup is slightly more likely to yield reward than left (i). However, he also observes advice from a confederate who *knows* which cup contains reward, and who he has learnt is very likely to give him misleading advice (ii). The confederate advises him to choose the right cup. (b) Combining these two probabilities convinces the subject that the left cup is more likely to be rewarding (iii). He chooses this cup, and discovers that he has won on this trial. (c) This yields dissociable prediction errors in three distinct frames of reference; firstly, in the traditional frame of reference of reward on one's own actions; secondly, in the frame of reference of which cup is likely to yield reward; and thirdly, in the

frame of reference of the *intentions* of the other player – how likely he is to be helpful or otherwise in the future.

References

1. **Baron-Cohen S, Leslie AM, Frith U (1985) Does the autistic child have a "theory of mind"? Cognition 21:37-46.**
2. **Baron-Cohen S, Ring HA, Wheelright S, Bullmore ET, Brammer MJ, Simmons A, Williams SCR (1999) Social intelligence in the normal and autistic brain: an fMRI study. Eur J Neurosci 11:1891-1898.**
3. **Becerra L, Breiter HC, Wise R, Gonzalez RG, Borsook D (2001) Reward circuitry activation by noxious thermal stimuli. Neuron 32:927-946.**
4. **Behrens TE, Hunt LT, Rushworth MF (2009) The computation of social behavior. Science 324:1160-1164.**
5. **Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MSF (2008) Associative learning of social value. Nature 456:245-249.**
6. **Brothers L (1990) The social brain: a project for integrating primate behavior and neurophysiology in a new domain. Concepts Neurosci 1:27-51.**
7. **Camerer CF (2003) Behavioral game therory: experiments in strategic interaction. Princeton: Princeton University Press.**
8. **Camerer CF, Ho T-H, Chong J-K (2004) A cognitive hierarchy model of games. Q J Econ 119:861-898.**
9. **Campbell-Meiklejohn DK, Bach DR, Roepstorff A, Dolan RJ, Frith C (2010) How the opinion of others affects our valuation of objects. Curr Biol 20:1165-1170.**
10. **Cohen YE, Andersen RA (2002) A common reference frame for movement plans in the posterior parietal cortex. Nat Rev Neurosci 3:553-562.**
11. **Coricelli G, Nagel R (2009) Neural correlates of depth of strategic reasoning in medial prefrontal cortex. Proc Natl Acad Sci USA 106:9163-9168.**
12. **de Quervain DJ, Fischbacher U, Treyer V, Schellhammer M, Schnyder U, Buck A, Fehr E (2004) The neural basis of altruistic punishment. Science 305:1254-1258.**
13. **Deaner RO, Khera AV, Platt ML (2005) Monkeys pay per view: adaptive valuation of social images by rhesus macaques. Curr Biol 15:543-548.**
14. **Delgado MR, Schotter A, Ozbay EY, Phelps EA (2008) Understanding overbidding: using the neural circuitry of reward to design economic auctions. Science 321:1849-1852.**
15. **Dunbar R (1993) Coevolution of neocortex size, group size and language in humans. Behav Brain Sci 16:681-735.**

16. Duncan J, Owen AM (2000) Common regions of the human frontal lobe recruited by diverse cognitive demands. Trends Neurosci 23:475-483.
17. Falk A, Fischbacher U (2006) A theory of reciprocity. Games and economic behavior 54:293-315.
18. Fehr E, Schmidt KM (1999) A theory of fairness, comptetition and cooperation. Q J Econ 114:817-868.
19. Fehr E, Fischbacher U (2003) The nature of human altruism. Nature 425:785-791.
20. Fletcher PC, Happé F, Frith U, Baker SC, Dolan RJ, Frackowiak RSJ, Frith C (1995) Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57:109-128.
21. Gallagher HL, Happé F, Brunswick N, Fletcher PC, Frith U, Frith C (2000) Reading the mind in cartoons and stories: an fMRI study of "theory of mind" in verbal and nonverbal tasks. Neuropsycholgia 38:11-21.
22. Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci USA 105:6741-6746.
23. Harbaugh WT, Mayr U, Burghart DR (2007) Neural responses to taxation and voluntary giving reveal motives for charitable donations. Science 316:1622-1625.
24. Hare TA, Camerer CF, Knoepfle DT, O'Doherty JP, Rangel A (2010) Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. J Neurosci 30:583-590.
25. Homans G (1958) Social behavior as exhange. Am J Sociol 62:597-606.
26. Izuma K, Saito DN, Sadato N (2008) Processing of social and monetary rewards in the human striatum. Neuron 58:284-294.
27. King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR (2005) Getting to know you: reputation and trust in a two-person economic exchange. Science 308:78-83.
28. King-Casas B, Sharp C, Lomax-Bream L, Lohrenz T, Fonagy P, Montague PR (2008) The rupture and repair of cooperation in borderline personality disorder. Science 321:806-810.
29. Klein JT, Deaner RO, Platt ML (2008) Neural correlates of social target value in macaque parietal cortex. Curr Biol 18:419-424.
30. Klucharev V, Hytonen K, Rijpkema M, Smidts A, Fernadez G (2009) Reinforcement learning signal predicts social conformity. Neuron 61:140-151.
31. Knoch D, Schneider F, Schunk D, Hohmann M, Fehr E (2009) Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. Proc Natl Acad Sci USA 106:20895-20899.
32. Knoch D, Gianotti LR, Pascual-Leone A, Treyer V, Regard M, Hohmann M, Brugger P (2006) Disruption of right prefrontal cortex by low-frequency repetitive transcranial magnetic stimulation induces risk-taking behavior. J Neurosci 26:6469-6472.

33. Moll J, Krueger F, Zahn R, Pardini M, de Oliveira-Souza R, Grafman J (2006) Human fronto-mesolimbic networks guide decisions about charitable donation. Proc Natl Acad Sci USA 103:15623-15628.
34. Nagel R (1995) Unravelling in guessing games: an experimental study. Am Econ Review 85:1313-1326.
35. Nitz DA (2009) Parietal cortex, navigation and the construction of arbitrary reference frames for spatial navigation. Neurobiol Learn Mem 91:179-185.
36. O'Doherty JP, Dayan P, Friston KJ, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38:329-337.
37. Poldrack RA (2006) Can cognitive processes be inferred from neuroimaging data? Trends Cogn Sci (Regul Ed) 10:59-63.
38. Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S (2004) The role of the medial frontal cortex in cognitive control. Science 306:443-447.
39. Rilling J, Gutman D, Zeh T, Pagnoni G, Berns G, Kilts C (2002) A neural basis for social cooperation. Neuron 35:395-405.
40. Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, Cohen JD (2004) The neural correlates of theory of mind within interpersonal interactions. Neuroimage 22:1694-1703.
41. Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. Science 300:1755-1758.
42. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593-1599.
43. Seymour B, Daw ND, Dayan P, Singer T, Dolan RJ (2007) Differential encoding of losses and gains in the human striatum. J Neurosci 27:4826-4831.
44. Singer T, Seymour B, O'Doherty JP, Stephan KE, Dolan RJ, Frith CD (2006) Empathic neural responses are modulated by the perceived fairness of others. Nature 439:466-469.
45. Skyrms B (2004) The stag hunt and evolution of social structure. Cambridge: Cambridge University Press.
46. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46:1004-1017.
47. Thaler RH (1988) Anomalies: the ultimatum game. J Econ Persp 2:195-206.
48. Tricomi E, Rangel A, Camerer CF, O'Doherty JP (2010) Neural evidence for inequality-averse social preferences. Nature 463:1089-1091.
49. van den Bos W, van Dijk E, Westenberg M, Rombouts SARB, Crone EA (2009) What motivates repayment? Neural correlates of reciprocity in the trust game. Soc Cog Aff Neurosci 4:294-304.
50. Van Overwalle F (2009) Social cognition and the brain: a meta-analysis. Human brain mapping 30:829-858.

51. Yamagishi T, Horita Y, Takagishi H, Shinada M, Tanida S, Cook KS (2009) The private rejection of unfair offers and emotional commitment. Proc Natl Acad Sci USA 106:11520-11523.
52. Yoshida W, Dolan RJ, Friston KJ (2008) Game theory of mind. PLoS Comp Biol 4:e1000254.
53. Yoshida W, Seymour B, Friston KJ, Dolan RJ (2010) Neural mechanisms of belief inference during cooperative games. J Neurosci 30:10744-10751.

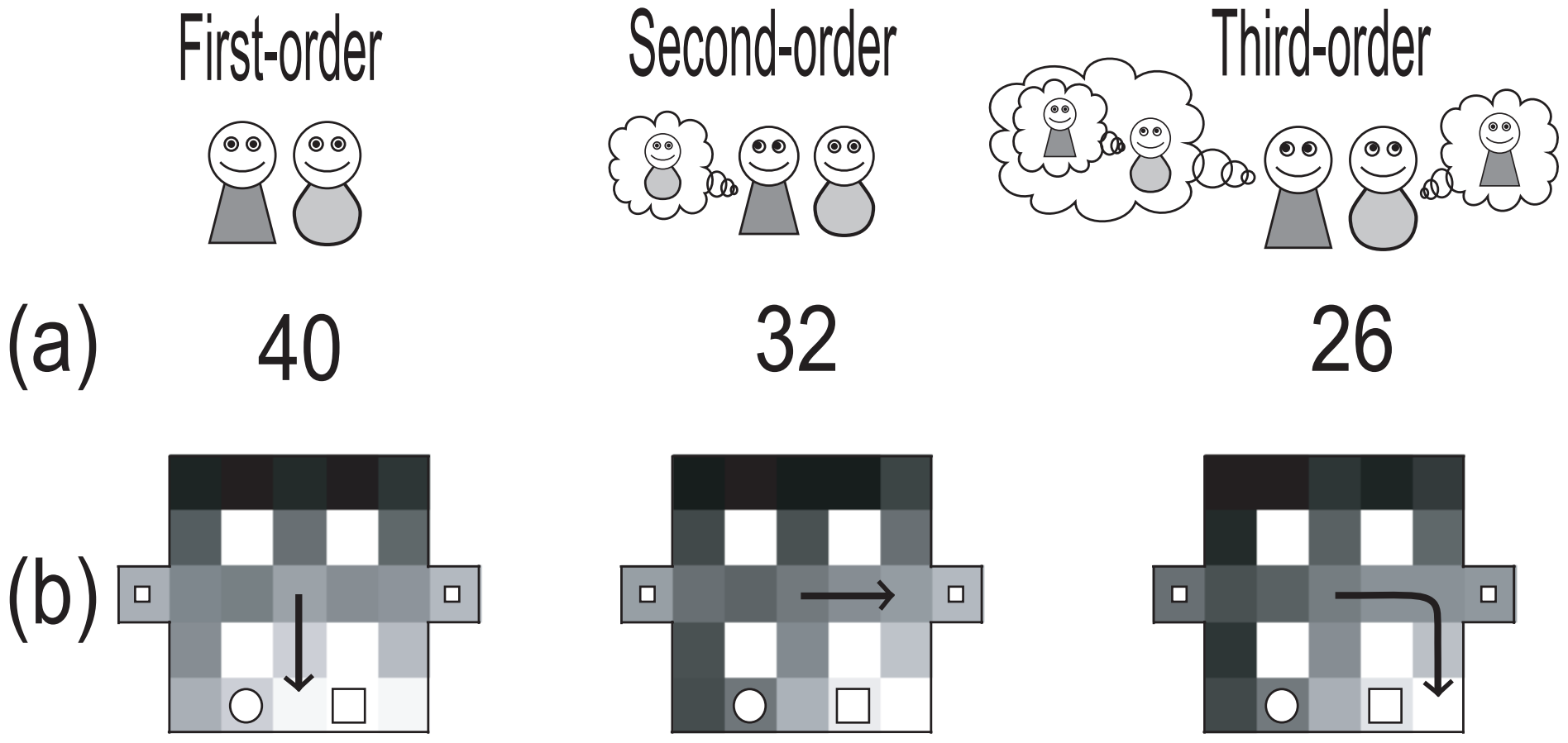

Figure1

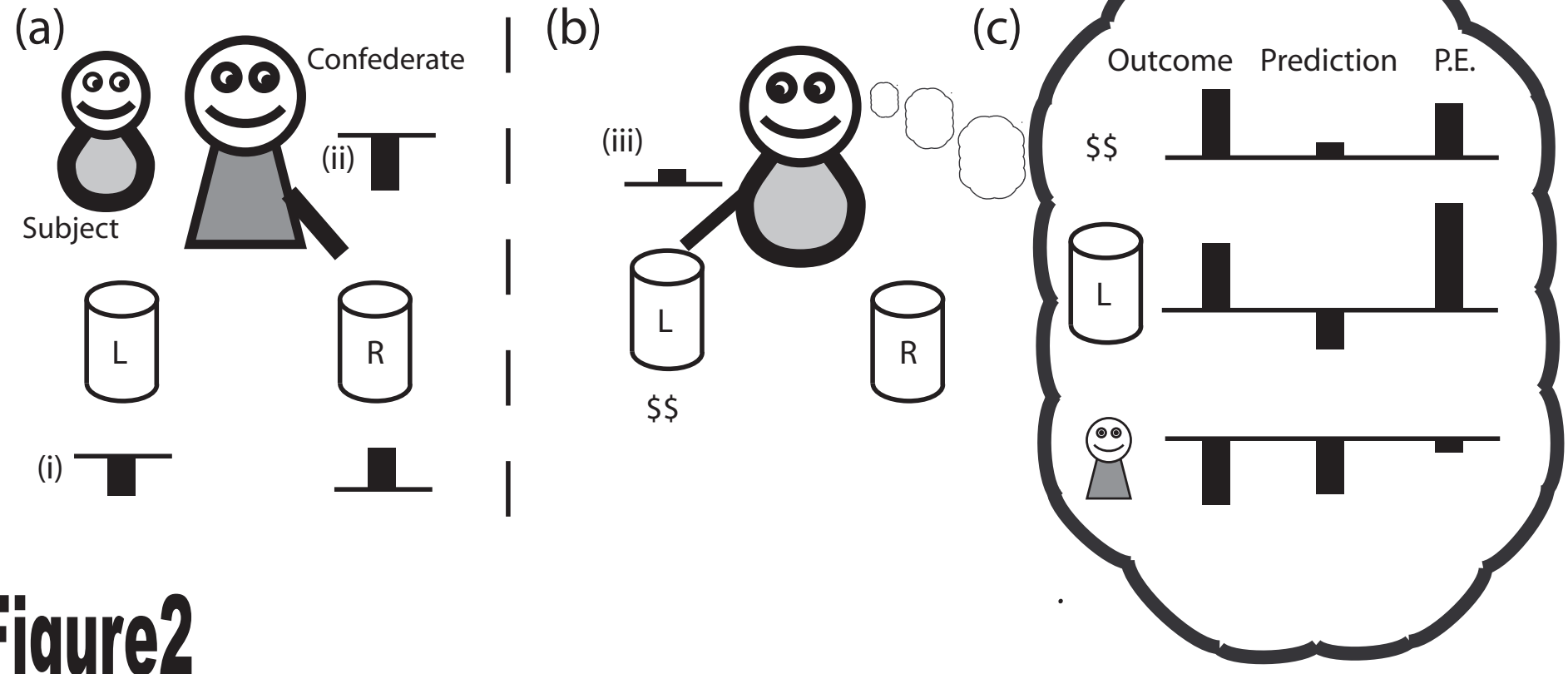
(a)
Confederate
(ii)
Subject
L
R
(i)
(b)
(iii)
L
R
$$
(c)
Outcome
Prediction
P.E.
$$
L

Figure2