frontiers in
BEHAVIORAL NEUROSCIENCE

HYPOTHESIS AND THEORY ARTICLE
published: 08 September 2009
doi: 10.3389/neuro.08.023.2009

Altruistic learning

***Ben Seymour*[1,2]*, *Wako Yoshida*[1] *and Ray Dolan*[1]**

[1] *Wellcome Trust Centre for Neuroimaging, University College London, London, UK*
[2] *ESRC Centre for Economic Learning and Social Evolution, University College London, London, UK*

Edited by:
Daeyeol Lee, Yale University, USA

Reviewed by:
Michael J. Frank, Brown University, USA
Daeyeol Lee, Yale University, USA

***Correspondence:**
Ben Seymour, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London WC1N 3BG, UK.
e-mail: bj.seymour@gmail.com

The origin of altruism remains one of the most enduring puzzles of human behaviour. Indeed, true altruism is often thought either not to exist, or to arise merely as a miscalculation of otherwise selfish behaviour. In this paper, we argue that altruism emerges directly from the way in which distinct human decision-making systems learn about rewards. Using insights provided by neurobiological accounts of human decision-making, we suggest that reinforcement learning in game-theoretic social interactions (habitisation over either individuals or games) and observational learning (either imitative of inference based) lead to altruistic behaviour. This arises not only as a result of computational efficiency in the face of processing complexity, but as a direct consequence of optimal inference in the face of uncertainty. Critically, we argue that the fact that evolutionary pressure acts not over the object of learning ('what' is learned), but over the learning systems themselves ('how' things are learned), enables the evolution of altruism despite the direct threat posed by free-riders.

Keywords: reinforcement learning, altruism, evolution, neuroeconomics, strong reciprocity, theory of mind, free-rider problem

INTRODUCTION

Many social interactions are self-beneficial if we behave positively and pro-cooperatively towards others. Opportunities to benefit from cooperation are widespread, and reflect the extrinsic fact that the natural environment is often best harvested, insofar as rewards can be accrued and threats avoided, by working together. But the decision to cooperate is not always straightforward, as in some situations it leaves us vulnerable to exploitation by others.

Game theory specifies a set of potential social interactions in which outcomes of cooperation and defection systematically differ, allowing both experimentalists and theoreticians to probe an individual's propensity for cooperation in different situations (Camerer, 2003). These outcomes typically vary in the extent to which competitive actions may seem preferable and where a short-sighted temptation to exploit the cooperativeness of others has a capacity to subvert cooperation later. Fortunately, the ability to look beyond the immediate returns of defection towards longer-term cooperation allows humans to escape from otherwise competitive equilibria, and this can be viewed as a hallmark of rational, sophisticated behaviour.

However, humans appear to behave positively towards each other in situations in which there is no capacity to benefit from long-term cooperation: for instance, when they play single games in which they never meet the same opponent again, and when their identities are kept anonymous (Fehr et al., 1993; Berg et al., 1995; Fehr and Fischbacher, 2003). This removes the capacity for both direct reciprocity (tit-for-tat) (Trivers, 1971; Axelrod, 1984), and the ability to earn a cooperative and trustworthy reputation that can be communicated by a third-party (Harbaugh, 1998; Bateson et al., 2006; Ariely and Norton, 2007). Furthermore, they will do this even if it is costly to themselves (Xiao and Houser, 2005; Henrich et al., 2006). From an economic perspective this appears to be genuinely altruistic, being strictly irrational since it incurs a direct personal cost with no conceivable long-term benefit.

Arguments against altruistic interpretations of experimentally observed behaviour include suggestions that individuals do not understand the rules of the game, are prone to misbelieve they (or their kin) will interact with opponents again in the future, or falsely infer they are being secretly observed and accordingly act to preserve their reputation in the eyes of experimenters (Smith, 1976). However, the widespread observation of altruism (both rewarding and punishing) across cultures (Henrich et al., 2001), and within meticulously designed experiments conducted by behavioural economists provide compelling support for its presence as a clear behavioural disposition. Furthermore, in fMRI experiments, altruistic actions correlate with brain activity, suggesting that they derive from some sort of intended or motivated behaviour and are not an expression of mere 'effector noise' (i.e. decision error) (de Quervain et al., 2004).

The very existence of altruism raises the difficult question as to why evolution has allowed otherwise highly sophisticated brains to behave so selflessly. This directs attention towards the decision-making systems that subserve economic and social behaviour (Lee, 2006, 2008; Behrens et al., 2009), and questions whether they are structured in such a way that yields altruism either inadvertently, or necessarily. The broader consequence is that if they do, then this reframes the question regarding the ultimate (evolutionary) causes of altruism towards the evolution of these very decision systems, and away from the phenomenological reality of altruism *per se*.

In this paper, we first review the structure of distinct human decision-making systems by considering a goal-directed (cognitive) system, a habitual system, and an innate (Pavlovian) action system and their interactions. We consider how these systems might operate in social contexts where the key problem is how to make optimal decisions when outcomes depend on the uncertainty associated with other agents and their motives. In the face of such computational complexity, we then consider how optimal actions can

be approximated by habit-based decision-making when outcomes are reliably predicted. In this context – through habits – altruism emerges as a consequence of a net economy of computational cost. We also consider the problem of evaluating the best policy when the payoff matrix is unknown but where individuals have an opportunity to learn from others. Observational learning rests upon inferences that might utilise such conspicuous attributes as their personal wealth. We frame observation as an inverse reinforcement learning problem, and consider value functions (including goals and subgoals) that are inferred from others actions, as well as by simpler strategies such as imitation. Notably, with incomplete information – a consequence of not being around to observe the long-term benefits of pro-cooperative behaviours, altruistic outcomes may be inferred as surrogate goals. In this context, altruism arises through optimal inference with incomplete information.

THE ARCHITECTURE OF DECISION-MAKING

Studies of decision-making in behavioural neuroscience and psychology have tended to concentrate on elemental decision-making problems, such as reward accrual in simple, stochastic, non-social environments. This enterprise has been very successful and has combined ingenious experimental designs with more classical focal brain lesion paradigms to yield insights into the underlying structure of decision-making systems. One key emerging insight is the likelihood that there is no singly monolithic decision-making system in the brain. Indeed, the best evidence suggest there are at least three distinct decision-making systems comprising a goal-directed, habitual, and innate (Pavlovian) system – with behavioural control being an admixture of cooperation or independence (Dickinson and Balleine, 2002; Dayan, 2008).

Goal-directed decision-making systems function by building an internal model of the environment. In the simplest case this may simply involve representing the identity of the expected outcome. In more complicated instances, it involves detailed knowledge of the structure of the environment and one's position within it. Although a goal-directed system may subsume several distinct sub-mechanisms, a wide variety of evidence suggest it localises to prefrontal cortex (Daw et al., 2006; Kim et al., 2006; Valentin et al., 2007), hippocampus (Corbit and Balleine, 2000; Kumaran and Maguire, 2006; Lengyel and Dayan, 2007) and dorsomedial striatum (Balleine and Dickinson, 1998; Corbit et al., 2003; Yin et al., 2005).

Habits, on the other hand, lack specific knowledge of the outcome of their decisions. In the parlance of computer science their values are 'cached', and represent only a scalar quantity which describes how good or bad an action is (Daw et al., 2005). In animal learning, such values are characterised by their insensitivity to devaluation: changes in state (e.g. moving from hunger to satiety) do not alter the value of the action, since there is no access to the new value of the goal (Dickinson and Balleine, 1994; Daw et al., 2005). Habits are acquired through experience, and 'rationalised' on account of their reliability in predicting rewarding outcomes. This efficiency derives entirely from the way in which they learn: rewards reinforce actions that are statistically predictive of their occurrence, with reinforced actions acquiring value through simple associative learning rules (Rescorla and Wagner, 1972; Holman, 1975; Adams and Dickinson, 1981). These are well described by Reinforcement Learning algorithms (such as Q learning and SARSA; Sutton and Barto, 1998), and localise to dorsolateral striatum (O'Doherty et al., 2003; Tricomi et al., 2009) and dopaminergic projections from substantia nigra.

Control over decisions is often dynamic and frequently transfers from goal-directed mechanisms (early in a task) to a habit-based system (late in a task). Indeed, this transfer can be manipulated by selective lesions to the neural substrates that underlie each of these systems (Balleine et al., 2009). In formalising accounts of how these systems interact current views centre on the idea of control being mediated by the respective uncertainties with which each system predicts outcomes, a view that provides a reasonable normative account of experimental findings (Daw et al., 2005). At a broader level, the evolutionary rationale for such a dual system is based on computational cost, since habits are vastly less resource demanding than goal-directed mechanisms.

Lastly, animals including humans have an innate, 'hard-wired', decision system. This is often referred to as a Pavlovian system, characterised by the expression of values and responses acquired through simple state-based associative learning. Unconditioned and conditioned Pavlovian responses represent an evolutionarily acquired behavioural repertoire that reflect basic, reliable knowledge gleaned from an organisms evolutionary history: embodying such knowledge structures that approaching sweet tasting fruit and withdrawing from bitter tasting fruit are inherently useful responses to enact. But whereas, on average, this inbuilt knowledge structure is enormously valuable to a naïve individual, it may also be a curse in the (usually) uncommon situations in which it is incorrect. The competitive (inhibitory) interaction between decisions based on experience (instrumental habit and goal-directed mechanisms) and those based on Pavlovian impulse localises to brain regions such as the amygdala and ventral striatum (Cardinal et al., 2002; Seymour and Dolan, 2008). This interaction reflects the classic tension between apparently emotional irrational and rational cognitive systems whereby the emotional expresses an apparent irrationality by way of some peculiarity of the environment.

DECISION-MAKING IN GAMES

A challenge for decision neuroscience is to understand how basic decision-making systems operate within socially interactive environments. Consider the game in **Table 1**: the repeated Prisoner's

Table 1 | An example payoff matrix of two-player Prisoner's dilemma game in which each player can choose either to 'cooperate' or 'defect'. The Left-side numbers represent the payoffs for the first player and the right-side numbers represent the payoffs for the second. Payoffs are symmetric, and chosen so that the sum of the payoffs is greatest when both choose cooperate and least when both players choose defect. However, each player earns the most if he chooses to defect when the other cooperates. Thus, the unique subgame perfect Nash Equilibrium of this game is for both players to defect.

		Player 2	
		Cooperate	**Defect**
Player 1	**Cooperate**	10/10	0/15
	Defect	15/0	1/1

dilemma. Subjects must choose between one of two actions: cooperate or defect, and their payoff depends on this and the choice of the opponent. Now consider a goal-directed, cognitive decision-making policy in the game, which has the ability to consider multiple future hypothetical scenarios (**Figure 1A**). If you neither know, nor care, what the other player does then the best strategy is to defect on the first round, since the outcome is always better regardless of what the other player does. For the same reason, even if you know what he/she will do, it is still better to defect.

However, it is also clear that in the long run, both players are better off if they cooperate: this mutually prescribes the best exploitation of environmental resources. Clearly, you need some way of both knowing that your opponent is committed to cooperation as well as a means of signalling to him/her your intention to cooperate. That is, you need to know that she is sophisticated enough to realise that cooperation is worthwhile, and you yourself need to be sophisticated enough to realise this. There is nothing truly altruistic about this, since you are both just trying to maximise your own payoff in an environment that contains another intelligent agent.

Thus, the existence of another intelligent agent in the environment makes the problem more complex than simpler decision-making problems that exploit inanimate environments. In the latter, the payoff probability usually depends fully on the observable states (they are 'fully observable Markov decision problems'; Bellman, 1957). That is, although the payoff may be probabilistic (either involving risk or ambiguity or both), your predictions depend in no way on how you came to arrive at that state in the first place. In social interactions, this assumption does not apply because outcomes depend on what the state thinks about you. If you have recently behaved uncooperatively, then this history negatively influences the payoff you expect to receive. That is, the outcome depends on unobservable states in the environment (making the problem 'partially observable'). If you find yourself in a seemingly identical state to a previous occasion, for instance playing opponent *x* in the game *y*, then the expected payoffs are not independent of how you got there, since opponent *x* may have a memory of you.

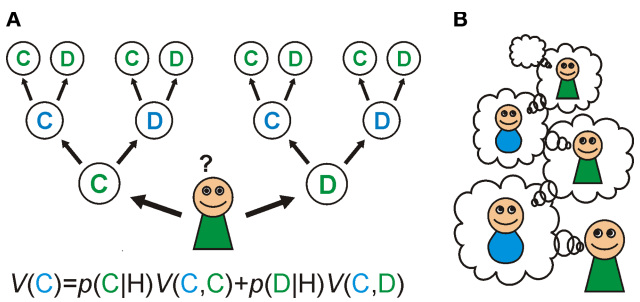

FIGURE 1 | Goal-directed learning of Prisoner's dilemma. (A) In goal-directed learning, players learn the probability of other player's action: cooperation (C) or defection (D) based on the history of their actions (H) as p(C|H) and p(D|H). They estimate the value of their own actions: V(C) and V(D) using the prediction from the learned model and the expected reward from the pair of actions. **(B)** In social games, the model of others leads a recursive process: my model of your action includes a model of your estimate of my actions, and so on. Cooperation in the Prisoner's dilemma depends on these recursive representations; since when I decide to cooperate this time, I must estimate that you are going to cooperate next time as you believe that I am going to cooperate with you.

Consequently, social decision-making benefits greatly from constructing some sort of internal model of the key aspects of the environment. In social games this model needs to capture the intentions of the other player (a component of 'Theory of Mind'). Indeed, your model should also include your opponent's estimate of your intentions: with this model, you can strategically plan to signal to your opponent your intention to cooperate, knowing that it will change their model of you (**Figure 1B**). Accordingly, they should then be more willing to cooperate with you, and you will both be better off in the long run.

It can be seen that this sort of model of others' intentions, and their model of your intentions, captures features of reciprocity, trust, and reputation formation. Indeed maintaining cooperation is in everyone's selfish interest in repeated games when the end of play is not in sight. It does, however, require players to be able to resist the short-term temptation to exploit this mutual reciprocity by the treachery of defection.

Of course, there is no reason why an internal representation of an other-agent's belief model need stop at a knowing the representation of your intentions in their mind. At the next level, it could include your understanding that they know that you know that they know your intentions, and so on. That there are infinite levels of embedded beliefs that make any perfect decision-policy intractable, has inspired models of strategic behaviour that either bound the upper limit of reciprocal beliefs (an example of 'bounded rationality') (Camerer et al., 2004a; Hampton et al., 2008), or estimate the level of reciprocal belief in their opponent directly (Yoshida et al., 2008).

Experimental evidence indicates that in repeated games with the same opponent, people reliably cooperate, as theory predicts. Critically, however, the theory predicts that people shouldn't cooperate towards the end of repeated exchanges, when they play people that they will never meet again and who can't communicate with others that can. The observation that people do cooperate in these situations suggests something is either incorrect about the goal-directed model, or as we suggest, other decision-making systems compete to bias behaviour.

HABITISATION

In simple environments, habits allow you to navigate towards goals and avoid harm with speed and computational efficiency. Habits operate by allowing recently experienced rewards to reinforce actions that are statistically predictive of them. If an outcome is reliably predicted by an action, then the value of that action becomes high. The action set available to an individual at any one time is elicited by the configuration of cues and contexts in the environment, which represents the current 'state'. Importantly, habits don't themselves have access to any specific representation of their outcome, they merely know their value on an ordinal value scale.

Now consider action control in social games. Imagine you are playing a selfish but sophisticated opponent in endless rounds of the Prisoner's dilemma. Early in the game, your model-based system has the ability to consider multiple future rounds of the game, in which mutual cooperation is evaluated as valuable, since you know your opponent also knows this. Accordingly, mutual cooperation is

rewarded as the game dictates. After a few rounds, actions associated with 'cooperate' begin to reliably predict rewarding outcomes, and so the habit learning system, operating concurrently with goal-orientated systems, acquires greater predictive certainty. As this accrues, control is transferred to the habit system, and the computational cost of considering multiple future rounds is relieved. In simple terms, cooperation becomes more 'automatic'.

The critical feature of this type of habit learning is what defines the state by which the habit can be elicited. In animal learning theory, this is termed the 'discriminative stimulus', and is typically experimentally determined by the presence of a cue (Mackintosh, 1983). However, the discriminative stimulus in social games is more complex, and in principle could be determined by the nature of the game being played (Prisoner's dilemma, stag-hunt and so on) or by the identity of the opponent. Below, we consider both possibilities:

Imagine that you ignore the identity of your opponent, and by good fortune play the prisoners dilemma with multiple cooperative opponents: i.e. you exist within a population of sophisticated cooperators (**Figure 2A**). Different types of social interaction will have distinct payoff matrices: some will benefit cooperation, others will not. If you know which game you are playing when you engage in an action, then if your action (e.g. to cooperate) is reliably rewarded it will be accessible to acquisition by a habit learning system that simply encodes that in a given game, cooperation or competition is reliably beneficial.

Indeed even if the payoff matrix is not known, for instance in a novel game in an uncertain environment, a reasonable strategy may be to play by trial and error. This entails exploring different actions and seeing what the outcome is, in which case actions can be reinforced directly by habit systems. Simulation studies demonstrate how readily cooperative equilibria can be reached by simple associative algorithms (such as Q learning) without any model-based control at all (Littman, 1994; Claus and Boutilier, 1998; Hu and Wellman, 2004).

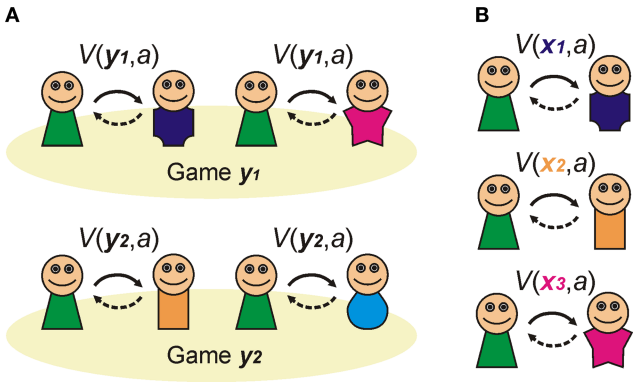

FIGURE 2 | Habitual learning of Prisoner's dilemma. (A) Habit learning in specific games. An agent plays an action *a* when in a particular state that is defined by the game type, e.g. game y1. If the outcome is rewarded, then the action is reinforced, and is more likely to be emitted when the same state is encountered again. **(B)** Habit learning with specific opponents. An agent (in green) plays action *a* (cooperate) when interacting with a particular agent (who defines the state, or discriminative stimulus S_D as x1, for example). If the outcome is rewarding, then the reward reinforces action *a*, such that it's value V(x1,a) increases, and is more likely to be chosen again in the same state in future.

Alternatively, you may choose to ignore the payoff matrix of the game, but concentrate instead on the identity of your opponent (**Figure 2B**). For instance, if you play a specific opponent in a variety of games, and she reliably cooperates with you to your benefit, then you may learn the habitual action to cooperate whenever you play her. In this way, she becomes a positive discriminative stimulus that evokes actions that engage pro-cooperatively with her.

The above mechanisms may acquire control of behaviour if several criteria are satisfied: the state and/or opponent are clearly discernable; the game (i.e. its payoff matrix) is relatively static (or changes slowly) allowing equilibria to be reached; and your internal preferences are stable. However, habit mechanisms are less reliable in the face of perceptual uncertainty, in which case an internal belief model of possible states may be required; if there are sudden changes in the environment that require rapid new learning, or a search for causal antecedents; or if your motivational state changes substantially (cooperation for food becomes less valuable when you are sated). Note that there is no evidence that habit systems 'switch off' in situations in which they behave poorly, rather their influence on control diminishes when their predictions become unreliable (Daw et al., 2005).

Although providing a plausible mechanism for social decision-making it turns out that, to date, evidence for habitised control of social behaviour is largely indirect. First, simple reinforcement learning algorithms do a remarkably good job at predicting behaviour in experiments across a variety of games (Erev and Roth, 1998, 2007). Second, neuroimaging studies show opponent-specific value-related responses accruing according to opponents' cooperativity/competitiveness in games (Singer et al., 2004). Third, neuroimaging studies have also identified dynamic reinforcement learning-like (prediction error) signals during games (King-Casas et al., 2005). Fourth, in single neuron recordings from non-human primates, lateral inter-parietal sulcus neurons in monkeys appear to encode value signals predicted by reinforcement learning in mixed-strategy games (Seo et al., 2009), which adds to previous observations that neurons in dorsolateral prefrontal and anterior cingulate cortex encode quantities related to choice and reinforcement history, respectively (Barraclough et al., 2004; Seo and Lee, 2008).

In reality, humans might be expected to habitise their actions in the context of state information that incorporates both opponent and game type. Although a diversity of subtly different payoff matrices may be common in experiments, it is likely that social interactions in different scenarios represent a relatively discrete set of payoff matrices. When there are small differences between different games, habit systems may generalise across salient features that have characteristic predictive power for beneficial outcomes.

OBSERVATIONAL LEARNING

One especially important social scenario arises when a person interacts with others who are significantly more expert at social interaction. This can occur for a number of reasons: if the payoff matrix that defines the interaction is unknown to us but known to others – either through their experience or private information; because information about other players is known to them but not to us – again through either experience or their own vicariously

acquired knowledge; of if they are more sophisticated – for instance they are more mature or intellectually able. In these situations, you have the choice to engage in interactions and acquire the information directly through your own experience or, better, to observe apparently successful social agents and vicariously acquire knowledge.

As long as success is discernable, as a hallmark of social expertise, then observational learning is likely to yield useful information. The computational problem becomes how to interpret the actions of others, and use observed actions to optimise your own. Computationally, *inverse* reinforcement learning describes this problem of how to reverse engineer observed actions to evaluate their values and goals, and is particularly difficult in situations in which actions do not immediately lead to their benefits. Unfortunately social interactions often display exactly this property: the benefits of cooperation are often long-term, through reputation formation and establishment of trust, and unless an observer has observational access to extended sequences of actions and their ultimate outcomes, the problem becomes even harder.

In general, there are two broad classes of solution. The first is simply to imitate others (Price and Boutilier, 2003). Imitation is the observational twin of habit learning, insofar as the resulting action has no specific representation of the outcome: it simply learns that a particular action is reliably performed in state s. The actions it bears are habit-like, elicited by a discriminative state that represents the environment in which they were learned. Accordingly, the ease of imitation depends on the discernability of the state of the observer. In **Figure 3A**, we illustrate this for a situation in which the state is defined by the game type: as long as it is clear to the subject that they are playing, say Game y = Prisoners Dilemma, then the imitated action will be 'cooperate when playing game y'. The imitated state-action pair could equally well be defined by the identity of the opponent. In this case, the resulting action will be 'cooperate when playing opponent x'. Note that the values of the actions can also be inferred by the frequency with which they are elicited by observation, allowing imitation to encode action values, and not just stimulus-responses.

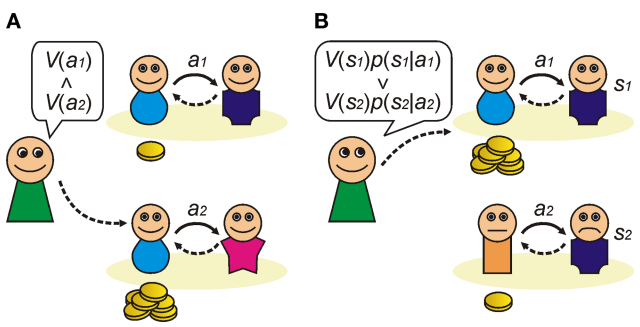

FIGURE 3 | Observation learning of Prisoner's dilemma. Observers learn the strategy from the observation of other players playing a game. **(A)** Imitation learning. An observer estimates the value of action *a* from other players' actions and simply imitates an action which maximises the payoff in a particular environment, which can be defined by game or opponent (or both). Here we show an example in which the environment is 'game = y' and it does not take into account the opponent's type (x) who they are playing with. **(B)** Inverse reinforcement learning. An observer estimates the *players'* value from their actions, for example using subgoals. This means the observer assumes that the players using a model-based learning; i.e. they have a forward model of their opponents. For example, in a repeated Prisoner's dilemma, cooperative actions (a_1) will predict a state of the other players' happiness (s_1) which leads mutual cooperation in the future. The value of action a is calculated as the value of state (e.g. other player's emotional state), $V(s)$, multiplied by the probability of occurrence of the state followed by the action, $p(s|a)$.

The second strategy is more complex, and involves trying to reverse engineer actions so as to evaluate their value or actual outcome (Ng and Russell, 2000). This requires constructing some sort of internal model of the action. For sequential actions, a computationally useful strategy is to represent subgoals – intermediate outcome states that appear to be reliable pre-requisites to eventual success (Abbeel and Ng, 2004). In the case of cooperative games, these subgoals ought to include the welfare of the other cooperators, since this is a powerful determinant of future cooperation. For example, in a repeated Prisoner's dilemma, sophisticated cooperators will themselves predict reward when their opponents cooperate with them, since they have a forward model of future beneficial interactions. Assuming their reward-predicting state is discernable by observations of their emotional state s (their happiness), then this state becomes a statistically reliable subgoal. That is, it follows that the inference that eliciting the state of happiness in another player is a valid predictor of an agent's success (**Figure 3B**).

Although in the case of the agent being observed this is merely an intermediary state in ultimately selfish reciprocal interactions, this information (and its selfishness) is not available to the observer. Even so, it is still valuable knowledge as long as the observer is fortunate enough to use the information in situations in which it actually is beneficial: i.e. in repeated social exchanges. As long as repeated social exchanges outnumber un-repeated exchanges, then observational inference is likely to be a better strategy than ignoring others.

Observational learning in games, and especially putative inverse reinforcement learning, remains relatively under-explored. It is well known that humans use both model-free (imitative) and model-based (inverse-inference) strategies when learning non-social actions through observation (Heyes and Dawson, 1990). Recent imaging evidence shows that people learn values through instruction using similar neural mechanisms involved in personal experience based learning (Behrens et al., 2008), and make inferences about values by pure third-party observation (Klucharev et al., 2009). Furthermore, pro-social feelings towards others (empathic reward), and it's neural representation, have been shown to be modulated by perceived similarity with that person (Mobbs et al., 2009), as one might predict from perspective-taking theories of social observation (Wolpert et al., 2003).

DISCUSSION

We have argued that consideration of the neurobiological mechanisms of learning and decision-making in humans can yield an explanatory account of true altruism. At the heart of this account are the learning systems that allow the brain to optimise reward and efficiency in complex environments. Critically, since evolution is likely to operate primarily over learning and decision mechanisms, and not the content of those systems – how they learn, not what they learn, the ensuing altruistic behaviours are perfectly permissible, despite the fact that they may in some instances become

strictly irrational. This is strengthened by the fact that habit-based and observational learning systems have uses way beyond social decision-making *per se*. The latter, for instance, is elegantly utilised in complex behaviours such as food preparation, tool use, and even language. Hence evolutionary selection for such mechanisms may be driven by a much broader range of decision-making problems than purely social interaction. Accordingly, such learning based accounts may offer both proximate and ultimate explanations for altruism.

The value of the inherent flexibility of learning systems is that it allows them to adapt to a wide range of potentially new and unexpected situations, appropriate for the diversity of the natural environment. But this flexibility carries the cost of inadvertently allowing individually economically disadvantageous actions to emerge, albeit rarely. However, we propose that on average these costs are heavily outweighed by benefits. Part of this supposition incorporates the fact that an innate representation of the caveats of flexible learning in social decision-making (for instance: don't cooperate in one-shot, anonymous exchanges in large groups) is itself cripplingly complex and maladaptive to novelty (it itself becomes a form of impulsivity). In other words, any social decision-making system that attempted to capture the enormous range of possible encounters and interactions, and individually specify optimal policies, would impair rather than augment decision-making under uncertainty. As such, efficient learning based systems are likely to be selected in the course of evolution.

Learning based accounts differ from the conventional approach of studying cooperation in behavioural economics, which often considers static, heuristic decision-policies, such as 'tit-for-tat', 'cooperate and punish', and 'free-ride'. Such models typically succumb to free-riders, including sophisticated (higher-order) free-riders that cooperate but don't enforce or encourage cooperation in others. However, a valuable insight of these models has been the recognition that resistance to free-riders can be provided by acquisition (and defence) of cultural norms of behaviour (Boyd and Richerson, 1988; Boyd et al., 2003; Bowles and Gintis, 2004). Key underlying components of norm-abidance are likely to be observational learning and inference based mechanisms, since these form simple elements of cultural learning. The current paucity of biologically implemented algorithmic models and mechanisms of observational and cultural learning is therefore likely to be an important area of future research. In particular, the relative privacy of culturally acquired information within specific groups is likely to be an important factor in the development of parochialism, which may further allow group-based selection of altruistic behaviour (Bernhard et al., 2006; Choi and Bowles, 2007).

Learning based accounts do not negate innate mechanisms of altruism in the brain. Such mechanisms are thought to underlie many aspects of human impulsivity and irrationality, through their occasionally inflexible competition with instrumental actions (Dayan et al., 2006). If cooperation was so consistently advantageous through human social evolution, that it is quite possible there might be some innate coding. Indeed, the environment in which the social brain evolved is likely to have had a much higher proportion of repeated interactions with the same individuals than our modern environment in which cooperation can occasionally be economically disadvantageous. Innate actions can be thought of as action priors over and above which more sophisticated goal-directed instrumental actions can assume control as experience accrues. Their Achilles heel, however, is the fact that they appear often difficult to overcome (inhibit) completely: they have a residual and significant weight that consistently biases actions in their favour. If such innate coding of cooperation exists in the human brain, then it follows that altruism would be akin to more basic forms of impulsivity.

We note that control by innate systems is characterised by the intrinsic (typically 'emotional') value of a stimulus, as well as by the action it elicits. Accordingly, the states associated with putatively pro-social innate actions could include that following the act of sharing, generosity or generation of equity (Tomasello et al., 2005). In this way, they become intrinsic internal rewards that, phenomenologically, are elicited because they are personally satisfying (and akin to non-social innate behaviours such as novelty-seeking (Wittmann et al., 2008)).

The complexity of different putative accounts of human altruism appeals to neuroscience as an arbitrator (Camerer et al., 2004b). Distinguishing different decision systems purely on anatomical grounds may be difficult, however: brain regions such as the striatum, orbitofrontal cortex, amygdala and hippocampus for instance, appear to be convergence areas for all decision systems. For example the observation of activation of striatum in a study on altruistic punishment (de Quervain et al., 2004), whilst providing a convincing illustration of the fact that such behaviour has a clear proximate basis, says little about the nature of that behaviour in terms of whether it is innate or learned. This underlines the importance for brain imaging techniques that have the ability to distinguish between competing models based on identifying coding of their underlying central parameters (O'Doherty et al., 2007), in situations in which behaviour alone is necessarily ambiguous (Yoshida et al., 2008).

Both habit-based and observation-based accounts of pro-social behaviour make specific experimental predictions. First, if the identities of others can act as discriminative stimuli, then cooperation should carry over between different games with the same individual. Second, if game types can act as discriminative stimuli, then cooperation should carry over between the same game with different individuals. Third, the duration of play should predict the degree of unfolding of cooperation towards the end of repeated games, since extended durations permit stronger habit formation and less susceptibility to anticipatory defection. Fourth, the operation of associative learning mechanisms should be determinable by the use of co-incident cues associated with previous cooperative or uncooperative players, which ought to bias individuals behaviour in future games: in fact evidence already exists for this (Vlaev and Chater, 2006; Chater et al., 2008). Fifth, observational learning can be studied directly by allowing individuals to passively watch interactions between others before engaging in similar games, or different games with the observed opponents. Indeed evidence does exist that previous observation has an influence on future social behaviour, in that people do seem to be biased towards the behaviour of others. What is more difficult to establish is exactly how this information is represented: either as a cached imitated value, or as a model-based representation.

Finally, we note that learning based accounts of altruism are by no means immune to exploitation by selfish and intelligent

learning agents. Any sophisticated model of other agents' behaviour can incorporate the fact that they are habit and observational learners. Consequently, highly sophisticated models of other agents could in theory incorporate representations of their different decision systems: thus knowing that people are habit learners gives predictive insight into what is likely to guide their behaviour in various situations. Whereas determining this might not always be simple to an agent from passive observation, it might be in part revealed by probing: intentionally behaving in a certain way (such as maliciously cultivating pro-social cultures) to manipulate how values are acquired by others, so that they can be exploited later.

REFERENCES

Abbeel, P., and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In Proceedings of the Twenty-first International Conference on Machine Learning, Banff, Alberta, Canada. ACM International Conference Proceeding Series. New York, NY, ACM.

Adams, C. D., and Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Q. J. Exp. Psychol. B Comp. Physiol. Psychol.* 33, 109–121.

Ariely, D., and Norton, M. I. (2007). Psychology and experimental economics: a gap in abstraction. *Curr. Dir. Psychol. Sci.* 16, 336–339.

Axelrod, R. (1984). The Evolution of Cooperation. New York, Basic Books.

Balleine, B. W., and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37, 407–419.

Balleine, B. W., Liljeholm, M., and Ostlund, S. B. (2009). The integrative function of the basal ganglia in instrumental conditioning. *Behav. Brain Res.* 199, 43–52.

Barraclough, D. J., Conroy, M. L., and Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nat. Neurosci.* 7, 404–410.

Bateson, M., Nettle, D., and Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biol. Lett.* 2, 412–414.

Behrens, T. E., Hunt, L. T., and Rushworth, M. F. (2009). The computation of social behavior. *Science* 324, 1160–1164.

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

Bellman, R. (1957). Dynamic Programming. Princeton, NJ, Princeton University Press.

Berg, J. E., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games Econ. Behav.* 10, 122–142.

Bernhard, H., Fischbacher, U., and Fehr, E. (2006). Parochial altruism in humans. *Nature* 442, 912–915.

Bowles, S., and Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theor. Popul. Biol.* 65, 17–28.

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. J. (2003). The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. USA* 100, 3531–3535.

Boyd, R., and Richerson, P. J. (1988). The evolution of reciprocity in sizable groups. *J. Theor. Biol.* 132, 337–356.

Camerer, C. F. (2003). Behavioural Game Theory: Experiments in Strategic Interaction. Princeton, NJ, Princeton University Press.

Camerer, C. F., Ho, T. H., and Chong, J. K. (2004a). A cognitive hierarchy model of games. *Q. J. Econ.* 119, 861–898.

Camerer, C. F., Loewenstein, G., and Prelec, D. (2004b). Neuroeconomics: why economics needs brains. *Scand. J. Econ.* 106, 555–579.

Cardinal, R. N., Parkinson, J. A., Hall, J., and Everitt, B. J. (2002). Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neurosci. Biobehav. Rev.* 26, 321–352.

Chater, N., Vlaev, I., and Grinberg, M. (2008). A new consequence of Simpson's paradox: stable cooperation in one-shot prisoner's dilemma from populations of individualistic learners. *J. Exp. Psychol. Gen.* 137, 403–421.

Choi, J. K., and Bowles, S. (2007). The coevolution of parochial altruism and war. *Science* 318, 636–640.

Claus, C., and Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In Proceedings of the Fifteenth National Conference on Artificial Intelligence. Menlo Park, CA, American Association for Artificial Intelligence, pp. 746–752.

Corbit, L. H., and Balleine, B. W. (2000). The role of the hippocampus in instrumental conditioning. *J. Neurosci.* 20, 4233–4239.

Corbit, L. H., Muir, J. L., and Balleine, B. W. (2003). Lesions of mediodorsal thalamus and anterior thalamic nuclei produce dissociable effects on instrumental conditioning in rats. *Eur. J. Neurosci.* 18, 1286–1294.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711.

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

Dayan, P. (2008). The role of value systems in decision making. In Better than Conscious? Decision Making, the Human Mind, and Implications for Institutions, C. Engel and W. Singer, eds (Frankfurt, MIT Press), pp. 51–70.

Dayan, P., Niv, Y., Seymour, B., and Daw, D. (2006). The misbehavior of value and the discipline of the will. *Neural Netw.* 19, 1153–1160.

de Quervain, D. J. F., Fischbacher, U., Treyer, V., Schelthammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004). The neural basis of altruistic punishment. *Science* 305, 1254–1258.

Dickinson, A., and Balleine, B. (1994). Motivational control of goal-directed action. *Anim. Learn. Behav.* 22, 1–18.

Dickinson, A., and Balleine, B. W. (2002). The role of learning in the operation of motivational systems. In Stevens' Handbook of Experimental Psychology, 3rd Edn., Vol. 3, Learning, Motivation, and Emotion, H. Pashler and R. Gallistel, eds (New York, John Wiley & Sons).

Erev, I., and Roth, A. E. (1998). Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* 88, 848–881.

Erev, I., and Roth, A. E. (2007). Multi-agent learning and the descriptive value of simple models. *Artif. Intell.* 171, 423–428.

Fehr, E., and Fischbacher, U. (2003). The nature of human altruism. *Nature* 425, 785–791.

Fehr, E., Kirchsteiger, A., and Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Q. J. Econ.* 108, 437–459.

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

Harbaugh, W. T. (1998). The prestige motive for making charitable transfers. *Am. Econ. Rev.* 88, 277–282.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001). In search of Homo economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* 91, 73–78.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., and Ziker, J. (2006). Costly punishment across human societies. *Science* 312, 1767–1770.

Heyes, C. M., and Dawson, G. R. (1990). A demonstration of observational-learning in rats using a bidirectional control. *Q. J. Exp. Psychol. B Comp. Physiol. Psychol.* 42, 59–71.

Holman, E. W. (1975). Some conditions for dissociation of consummatory and instrumental behavior in rats. *Learn. Motiv.* 6, 358–366.

Hu, J. L., and Wellman, M. P. (2004). Nash Q-learning for general-sum stochastic games. *J. Mach. Learn. Res.* 4, 1039–1069.

Kim, H., Shimojo, S., and O'Doherty, J. P. (2006). Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biol.* 4, e233. doi: 10.1371/journal.pbio.0040233.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., and Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78–83.

Klucharev, V., Hytonen, K., Rijpkema, M., Smidts, A., and Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron* 61, 140–151.

Kumaran, D., and Maguire, E. A. (2006). The dynamics of hippocampal activation during encoding of overlapping sequences. *Neuron* 49, 617–629.

Lee, D. (2006). Neural basis of quasi-rational decision making. *Curr. Opin. Neurobiol.* 16, 191–198.

Lee, D. (2008). Game theory and neural basis of social decision making. *Nat. Neurosci.* 11, 404–409.

Lengyel, M., and Dayan, P. (2007). Hippocampal contributions to control: The third way. *Adv. Neural Inf. Process. Syst.* 20, 889–896.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In Proceedings of the Eleventh International Conference on Machine Learning. San Francisco, CA, Morgan Kaufmann, pp. 157–163.

Mackintosh, N. J. (1983). Conditioning and associative learning. New York, Oxford University Press.

Mobbs, D., Yu, R. J., Meyer, M., Passamonti, L., Seymour, B.,

Calder, A. J., Schweizer, S., Frith, C. D., and Dalgleish, T. (2009). A key role for similarity in vicarious reward. *Science* 324, 900.

Ng, Y. N., and Russell, S. (2000). Algorithms for inverse reinforcement learning. In Proceedings of the Seventeenth International Conference on Machine Learning. San Francisco, CA, Morgan Kaufmann, pp. 663–670.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron* 38, 329–337.

O'Doherty, J. P., Hampton, A., and Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Ann. NY Acad. Sci.* 1104, 35–53.

Price, B., and Boutilier, C. (2003). Accelerating reinforcement learning through implicit imitation. *J. Artif. Intell. Res.* 19, 569–629.

Rescorla, R. A., and Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In Classical Conditioning, Vol. II, A. H. Black and W. F. Prokasy, eds (New York, Appleton-Century-Crofts).

Seo, H., Barraclough, D. J., and Lee, D. (2009). Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. *J. Neurosci.* 29, 7278–7289.

Seo, H., and Lee, D. (2008). Cortical mechanisms for reinforcement learning in competitive games. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 363, 3845–3857.

Seymour, B., and Dolan, R. (2008). Emotion, decision making, and the amygdala. *Neuron* 58, 662–671.

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., and Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron* 41, 653–662.

Smith, A. (1976). The Theory of Moral Sentiments, D. D. Raphael and A. L. Macfie, eds (Oxford, Oxford University Press).

Sutton, R. S., and Barto, A. G. (1998). Reinforcement Learning: An Introduction. Cambridge, MA, MIT Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behav. Brain Sci.* 28, 675–691.

Tricomi, E., Balleine, B. W., and O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *Eur. J. Neurosci.* 29, 2225–2232.

Trivers, R. (1971). The evolution of reciprocal altruism. *Q. Rev. Biol.* 46, 35–57.

Valentin, V. V., Dickinson, A., and O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *J. Neurosci.* 27, 4019–4026.

Vlaev, I., and Chater, N. (2006). Game relativity: how context influences strategic decision making. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 131–149.

Wittmann, B. C., Daw, N. D., Seymour, B., and Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron* 58, 967–973.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 358, 593–602.

Xiao, E., and Houser, D. (2005). Emotion expression in human punishment behavior. *Proc. Natl. Acad. Sci. USA* 102, 7398–7401.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., and Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *Eur. J. Neurosci.* 22, 513–523.

Yoshida, W., Dolan, R. J., and Friston, K. J. (2008). Game theory of mind. *PLoS Comput. Biol.* 4(12), e1000254. doi: 10.1371/journal.pcbi.1000254.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 June 2009; paper pending published: 13 July 2009; accepted: 14 August 2009; published online: 08 September 2009.

*Citation: Seymour B, Yoshida W and Dolan R (2009) Altruistic learning. Front. Behav. Neurosci. **3**:23. doi: 10.3389/neuro.08.023.2009*