nature neuroscience

© 2005 Nature Publishing Group http://www.nature.com/natureneuroscience

Perceptions of moral character modulate the neural systems of reward during the trust game

M R Delgado[1], R H Frank[2] & E A Phelps[1,3]

Studies of reward learning have implicated the striatum as part of a neural circuit that guides and adjusts future behavior on the basis of reward feedback. Here we investigate whether prior social and moral information about potential trading partners affects this neural circuitry. Participants made risky choices about whether to trust hypothetical trading partners after having read vivid descriptions of life events indicating praiseworthy, neutral or suspect moral character. Despite equivalent reinforcement rates for all partners, participants were persistently more likely to make risky choices with the 'good' partner. As expected from previous studies, activation of the caudate nucleus differentiated between positive and negative feedback, but only for the 'neutral' partner. Notably, it did not do so for the 'good' partner and did so only weakly for the 'bad' partner, suggesting that prior social and moral perceptions can diminish reliance on feedback mechanisms in the neural circuitry of trial-and-error reward learning.

The human striatum has been implicated as a critical structure in trial-and-error feedback processing and reward learning. In particular, the caudate nucleus, a structure linked to learning and memory in both animals[1–3] and humans[4–6], has been shown to have a role in processing affective feedback[7,8], with responses in this region varying according to properties such as valence and magnitude[9,10]. It has been shown that activation in the human caudate nucleus is modulated as a function of trial-and-error learning with feedback[5,6,11,12]. Activation in this region in response to reward feedback diminishes as, over time, cues begin to predict the correct action and outcome, thus making feedback less informative[11]. This has led to the hypothesis that the caudate nucleus may serve as a key component of an 'actor-critic' model processing the contingent behavior that led to the feedback, with the purpose of guiding future actions[13–15]. Recently, this role for the caudate in feedback processing has extended to social interactions in the economic domain using a repeated-interaction 'trust' game in which participants learned, through trial and error, whether their partners were trustworthy. As expected, activity in the caudate decreased over trials as feedback from the partners became more predictable[16].

Our study is motivated by the observation that not all choices are made on the basis of trial-and-error learning involving material rewards. Moral beliefs, for example, have been shown to influence economic choices. In one study, participants were more likely to cooperate in one-shot prisoner's dilemmas if they perceived their partners as cooperative after a brief period of informal conversation[17]. There is also evidence that many individuals are willing to work for lower salaries if they believe their employer's mission is morally praiseworthy[18].

To examine the influence of social and moral learning on choice, we used an iterated version of the trust game involving two-person interactions in which mutually beneficial outcomes are more likely if partners are trustworthy and perceive one another as such[19,20] (**Fig. 1a**). Participants could either keep $1 on a given trial or transfer it to a partner, in which case the partner would receive $3. The partner could either keep the entire $3 or give half of it back ('sharing'). In previous experiments involving economic games, participants learn which partners are trustworthy through trial and error[16,21,22]. In our version of the game, participants were instructed that they would be playing with three fictional partners and were given detailed, vivid descriptions of each partner's life events that indicated either praiseworthy, neutral or suspect moral character. Participants were told that the partners' responses might or might not be consistent with the descriptions given. Presentation of the trust trials were intermixed with 'lottery' trials, in which participants merely decided whether they wanted to play or not play a lottery for a chance to earn $1.50 reward.

Previous investigations of the reward circuitry and error-based learning[15,18,23,24] suggest three possible hypotheses. One is that no differences will be observed because rational economic decision makers will view information about the moral characteristics of partners as uninformative. Alternatively, information about moral characteristics might create expectations, and failed predictions that are based on those expectations might lead to an increase in the magnitude of responses in the brain circuitry associated with trial-and-error learning. A third possibility is that the bias induced by information about moral characteristics will modulate the brain mechanisms associated with trial-and-error learning, making participants less reliant on feedback or willing to discount such information. The present results suggest that moral beliefs can affect economic decision making, as a result of modulations in the human caudate nucleus that may influence the adjustment of choices based on trial-and-error feedback.

[1]Department of Psychology and [3]Center for Neural Science, 6 Washington Place, New York University, New York, New York 10003, USA. [2]Johnson School of Management, Cornell University, Ithaca, New York 14853, USA. Correspondence should be addressed to E.A.P. (liz.phelps@nyu.edu).

Received 25 July; accepted 21 September; published online 16 October 2005; doi:10.1038/nn1575

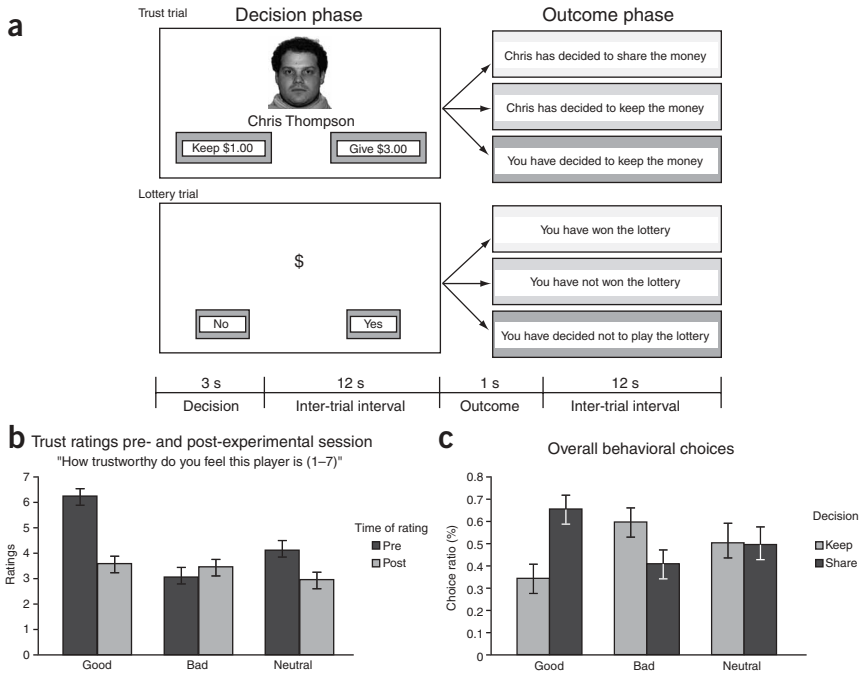

Figure 1 Experimental design and behavioral results. (**a**) A trial in the game was divided into a decision phase and an outcome phase and was played with one of three possible partners (good, bad or neutral) or against the computer (lottery). (**b**) Trustworthiness ratings, pre- and post-experiment (± s.e.m.). Participants' perception of moral character influenced trust ratings before experiment, which was adjusted after trial and error. (**c**) Behavioral choices during experimental sessions (± s.e.m.). Overall choices suggest participants were affected by the moral character of the fictional partner: rates of cooperation or sharing were higher when playing with the good partner, but not with the other partners.

© 2005 Nature Publishing Group http://www.nature.com/natureneuroscience

RESULTS

Behavioral results

After reading each fictional partner's biography, 12 participants gave trustworthiness ratings before each experimental session. These ratings confirmed that the manipulation affected the perceptions of trustworthiness (**Fig. 1b**). This was also reflected in choices: participants made more 'share' (that is, transfer of money) than 'keep' decisions when playing with the good partner ($t_{11} = 2.46$, $P < 0.05$); however, no such difference was observed when participants played with the bad ($t_{11} = -1.42$, $P = 0.18$) or neutral ($t_{11} = -0.05$, $P = 0.96$) partners (**Fig. 1c**). In addition, when comparing share decisions between partners (good versus bad), participants made more share decisions overall when playing with the good partner than with the bad ($t_{11} = 3.26$, $P < 0.01$) or neutral ($t_{11} = 2.0$, $P = 0.07$) partners. Finally, using reaction time data acquired during the experimental session, we observed that participants were faster to share when playing with the good partner compared to the bad ($t_{11} = -3.73$, $P < 0.005$) and neutral partners ($t_{11} = -1.89$, $P = 0.08$). Thus, biasing the participants according to the perceived moral characteristics of partners influenced decision-making overall.

As the experiment progressed, however, participants learned in some ways that the three partners were not responding quite in the manner suggested by their descriptions. For example, trustworthiness ratings acquired post-scanning differed from those acquired before scanning (**Fig. 1b**), as indicated in a two-way analysis of variance (ANOVA) by an interaction between partner and time of rating ($F_{2,22} = 11.78$, $P < 0.0001$). Further, after the experiment, we asked participants what percentage of time they believed each partner shared with them (good: mean = 44.58, s.d. = 22.10; bad: mean = 35.00, s.d. = 18.71; neutral: mean = 38.33, s.d. = 19.11). Although the mean scores ordered according to acquired bias (that is, good > neutral > bad), no significant differences were observed when comparing the two extreme partners (good versus bad: $t_{11} = 1.01$, $P = 0.34$). However, in spite of this explicit knowledge of similar response patterns, an investigation of early (first two of eight runs, containing roughly six trials per partner) and late decisions (last two runs) showed that participants made more share decisions when playing with the good partner than with the bad partner, on both early ($t_{11} = 3.99$, $P < 0.005$) and late ($t_{11} = 2.18$, $P < 0.05$) trials (see **Supplementary Fig. 1**). This suggests that explicit knowledge of feedback probabilities may not be completely predictive of choice behavior for partners of different perceived moral character.

To account for any asymmetry or special characteristics in partner biographies that may influence the results, we also conducted a separate trust game experiment with a separate set of participants and a new set of biographies, also depicted as morally good, bad and neutral (see **Supplementary Note**). This additional behavioral study replicated the previous results (**Supplementary Fig. 2**), further suggesting that moral and social characteristics can influence decision-making in an economic game.

Neuroimaging results

We conducted analyses of functional magnetic resonance imaging (fMRI) data separately for the outcome and decision phases. During

Table 1 Brain areas activated during the outcome phase

Region of activation	Voxels	Laterality	Talairach coordinates x	y	z
Positive > negative feedback					
Medial frontal gyrus (6)	17	L	–24	–5	59
Medial frontal gyrus (6)	15	R	13	8	1
Precuneus (7)	116	R	21	–79	44
Middle frontal gyrus (10/46)	43	L	–42	36	5
Caudate	222	R	15	20	7
Putamen	32	R	22	17	2
Caudate	**707**	**R**	**13**	**8**	**1**
Ventral striatum	12	R	16	0	–7
Inferior temporal gyrus (19)	14	R	55	–65	–2
Orbitofrontal gyrus (11)	11	R	29	27	–11
Parahippocampus gyrus	11	L	–42	–34	–16
Negative > positive feedback					
Insular cortex	14	R	54	8	13

The contrast of positive versus negative feedback yielded several regions defined by strength of effect ($P < 0.001$) and size (10 or more voxels). The reverse contrast yielded activation in one region. Brodmann's areas are depicted in parentheses (under the Region of activation column) when applicable. Cluster size (number of voxels) and laterality (right or left hemisphere) are also given. The stereotaxic coordinates of the peak of the activation are given according to Talairach space. The region of interest discussed in the paper is in bold.

Table 2 Brain areas activated during the decision phase

Region of activation	Voxels	Laterality	Talairach coordinates x	y	z
Share > keep decisions					
Inferior parietal cortex (40)	32	L	–66	–34	26
Insular cortex	43	L	–47	18	10
Lingual gyrus (18)	15	R	6	–70	–13
Putamen	112	L	–29	–16	3
Inferior occipital gyrus (18)	26	L	–36	–79	–2
Ventral striatum	**26**	**R**	**21**	**6**	**–4**
Fusiform gyrus (19)	34	R	12	–64	–7
Fusiform gyrus (19)	27	R	22	–67	–11
Keep > share decisions					
Perirhinal cortex	20	L	–35	1	–25
Perirhinal cortex	45	R	32	–6	–28

The contrast of share versus keep decisions yielded several regions defined by strength of effect ($P < 0.001$) and size (10 or more voxels). The reverse contrast yielded activation in a few regions. Specific annotations in the table have been previously described. The region of interest discussed in the paper is in bold.

Table 3 Brain areas activated during the decision phase contrasting bias-incongruent choices

Region of activation	Voxels	Laterality	Talairach coordinates x	y	z
Precentral gyrus (4/6)	79	L	–29	–17	68
Medial frontal gyrus (6)	67	R	2	1	57
Precentral gyrus (4)	24	L	–50	–20	55
Precuneus (7)	57	R	7	–73	44
Cingulate cortex (32/6)	12	R	4	8	47
Cingulate cortex (32)	**35**	**L**	**–4**	**17**	**38**
Cuneus (18/19)	20	R	2	–84	32
Inferior parietal cortex (40)	21	R	49	–32	25
Inferior parietal cortex (40)	14	L	–49	–37	20
Middle occipital gyrus (19)	20	L	–21	–88	17
Insular/frontal cortex	17	L	–63	–6	10
Insular cortex	40	L	–58	5	7
Insular cortex	21	L	–47	14	3
Insular cortex	**69**	**R**	**49**	**20**	**2**
Cerebellum	119	L	–31	–67	–22
Cerebellum	64	R	6	–78	–25

The contrast of decisions that were incongruent with behavioral bias (share with bad partner and keep with good partner versus the alternative choices) yielded several regions defined by strength of effect ($P < 0.001$) and size (10 or more voxels). Specific annotations in the table have been previously described. Region of interests discussed in the paper are in bold.

the outcome phase, we conducted a random-effects general linear model (GLM) analysis using each condition (good, bad, neutral and lottery outcome trials) and associated outcome (positive and negative feedback) as predictors. We generated statistical maps contrasting positive and negative feedback. Positive feedback referred to a monetary gain of $1.50 because the partner either cooperated (that is, 'shared back' the money) or the participant hit the lottery; negative feedback referred to no monetary gain ($0) either because the partner defected (that is, kept the money) or the participant did not hit the lottery. During the decision phase, we conducted a random-effects GLM analysis using each condition (good, bad and neutral decision trials) and the associated decision (share and keep) as predictors (for lottery trials, play decisions were also used as a predictor). We generated statistical maps (**Tables 1–3**) contrasting share and keep decisions. From contrasts generated during the outcome and decision phases, regions of interest (ROIs) in the striatum were identified based on peak activity, and *a priori* analyses for individual conditions were performed on each ROI using mean beta weights for each predictor.

Outcome phase

During the outcome phase, the contrast between positive (monetary gain) and negative (monetary loss) feedback yielded activation in the striatum, which was most robust in the ventral caudate nucleus (**Fig. 2a** and **Table 1**). We defined an ROI centered on the peak of this caudate activation. We then conducted additional analyses to investigate the effects of social and moral bias on feedback processing within this striatum ROI. First, we conducted a repeated-measures ANOVA using participants' mean beta weights. This analysis probed an interaction between moral character (good and bad versus neutral) and outcome (positive versus negative feedback). A significant interaction was observed ($F_{1,10} = 5.13$, $P < 0.05$), suggesting that perceived moral character influenced the underlying neural activity involved in feedback processing (**Fig. 2b**).

To further explore the source of this interaction, we conducted two ANOVAs comparing the good and bad partner separately with the neutral partner. There was a significant interaction in the response to positive and negative feedback between the good and neutral partner ($F_{1,10} = 5.8$, $P < 0.05$; **Fig. 2c**). Although a similar pattern was observed for the bad partner, this interaction did not reach significance ($F_{1,8} = 0.49$, $P = 0.51$; **Fig. 2d**).

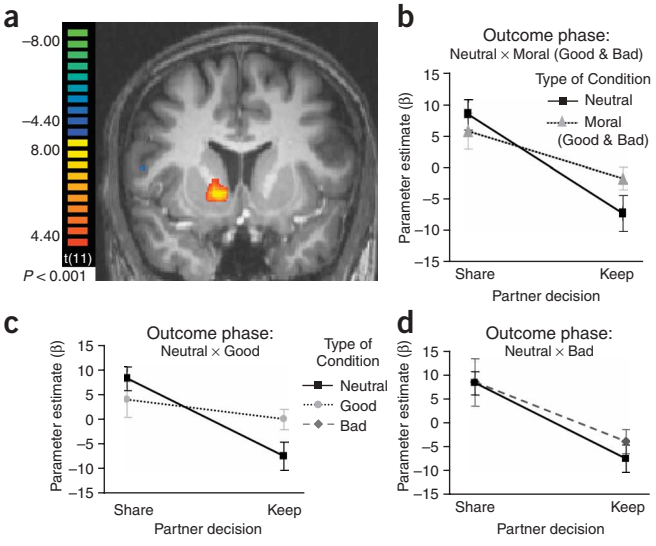

Figure 2 Caudate nucleus activation during outcome phase. (**a**) When contrasting positive and negative outcomes were collapsed across all conditions, activity was observed in the ventral portion of caudate nucleus ($x, y, z = 13, 8, 1$). (**b**) Repeated-measures ANOVAs using mean beta weights extracted from caudate nucleus ROI showed a significant interaction between moral character (good and bad versus neutral) and outcome (positive versus negative feedback). Separate ANOVAs investigated the source of the interaction by comparing the good and bad partner separately with the neutral partner. (**c**,**d**) A significant interaction between moral character and outcome was observed when comparing the neutral partner (**c**) with the good partner but (**d**) not with the bad.

© 2005 Nature Publishing Group http://www.nature.com/natureneuroscience

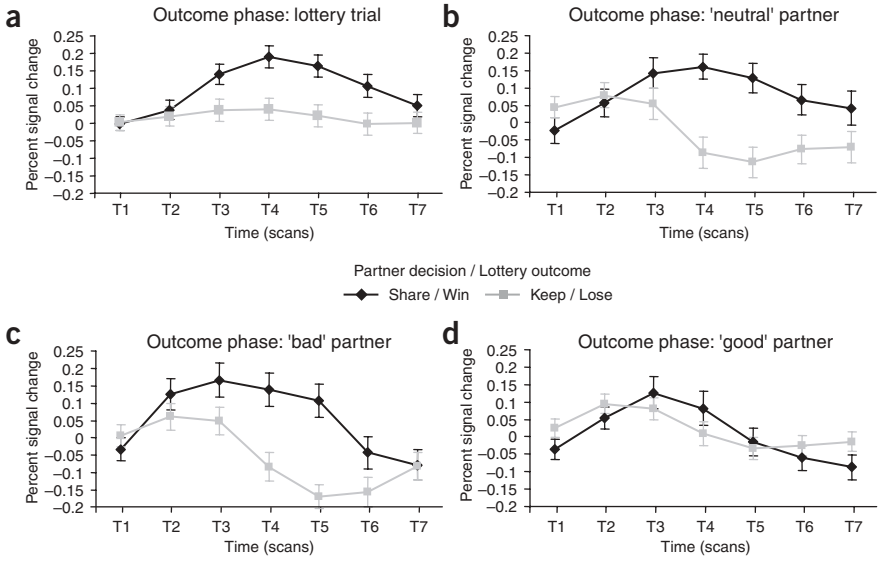

Figure 3 Time course of activation (± s.e.m.) of caudate nucleus ROI during the outcome phase for each condition. Time points are referred to as T1–T7 and correspond to 2 s each. (**a**–**c**) Time course data shows differential responses between positive and negative feedback during lottery trials, (**b**) during interactions with the neutral partner and (**c**) with the bad partner, to a lesser extent. (**d**) No differences were observed with the good partner.

Decision phase

During the decision phase, activation of the ventral striatum (nucleus accumbens and ventral putamen) was observed when contrasting share and keep decisions (**Table 2** and **Fig. 4**). We did not observe activity in the caudate nucleus, in agreement with the idea that although this area may be more involved with processing feedback to guide future behavior[13,14], ventral portions of the striatum may be more important for making predictions[14,26] and anticipating the outcomes of risky decisions[27,28]. We used mean beta weights extracted from this ROI in an ANOVA similar to the one described in the outcome phase to probe an interaction between moral character (good and bad versus neutral) and decision (share versus keep). No interaction was observed with this analysis ($F_{1,11} = 0.22$, $P = 0.65$). However, exploratory analysis of each individual condition suggested that some differences may exist. For example, when subjects were playing with the bad partner, mean beta weights for share decisions were significantly higher than those for keep decisions ($t_{11} = 2.64$, $P < 0.05$), and lottery play decisions ($t_{11} = 2.00$, $P < 0.05$). For the neutral partner, we observed a trend both when comparing beta weights for share and keep decisions ($t_{11} = 1.71$, $P = 0.06$) and when comparing them for share and lottery play decisions ($t_{11} = 1.70$, $P = 0.06$). We found no differences when participants were playing the good partner in either comparison (share × keep: $t_{10} = 1.29$, $P = 0.11$; share × lottery: $t_{11} = 0.40$, $P = 0.35$). Although the individual conditions analysis supports the idea that prior moral and social beliefs can influence the reward circuitry, the results must be interpreted cautiously because of the lack of overall interaction between moral character and decision.

Behavioral results showing that participants were more likely to share when playing with the good partner and keep when playing with the bad partner are consistent with the induced-bias manipulation. Given this, we posited that areas involved in cognitive control and

Additional post-hoc investigations of the mean beta weights in the ventral caudate ROI for each of the four individual conditions showed that positive outcomes were significantly different from negative outcomes for lottery ($t_{11} = 3.22$, $P < 0.005$) and neutral partner ($t_{10} = 5.03$, $P < 0.0005$) trials, consistent with previous studies showing this region differentiates between positive and negative outcomes[8,25] (**Fig. 3**). Less differentiation was observed for trials with the bad partner ($t_9 = 1.57$, $P = 0.08$). As expected from the interaction analysis, no significant differences were observed for trials involving the good partner ($t_{11} = 1.34$, $P = 0.11$). An additional analysis on percent signal change values showed that any difference observed between positive and negative feedback was maximal at 6–9 s after feedback delivery, consistent with previous studies of reward processing[8,25] (see **Supplementary Note**).

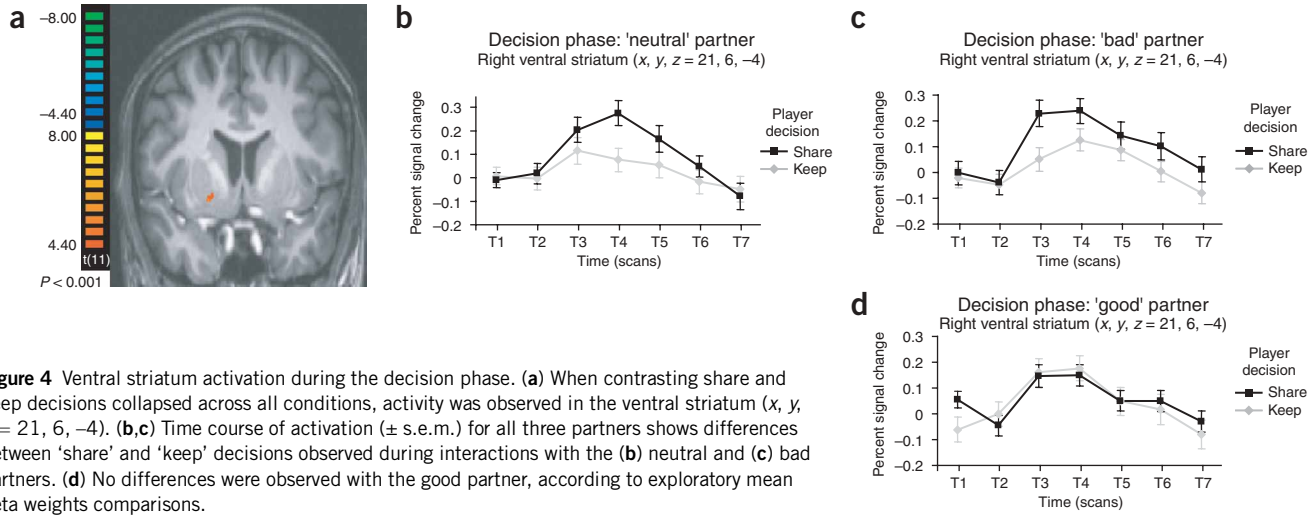

Figure 4 Ventral striatum activation during the decision phase. (**a**) When contrasting share and keep decisions collapsed across all conditions, activity was observed in the ventral striatum (*x*, *y*, *z* = 21, 6, −4). (**b**,**c**) Time course of activation (± s.e.m.) for all three partners shows differences between 'share' and 'keep' decisions observed during interactions with the (**b**) neutral and (**c**) bad partners. (**d**) No differences were observed with the good partner, according to exploratory mean beta weights comparisons.

© 2005 Nature Publishing Group http://www.nature.com/natureneuroscience

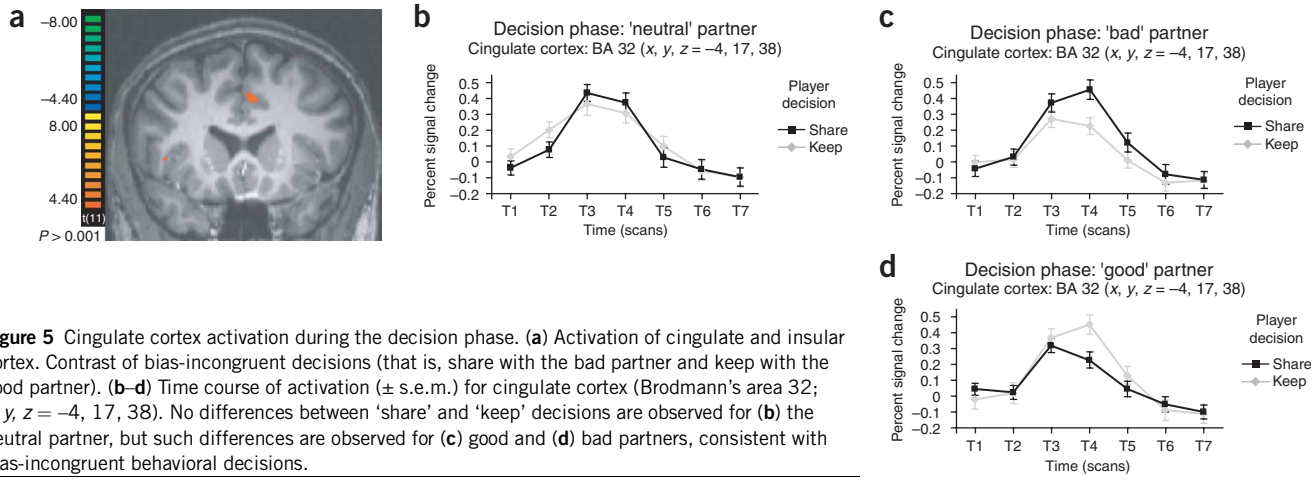

Figure 5 Cingulate cortex activation during the decision phase. (**a**) Activation of cingulate and insular cortex. Contrast of bias-incongruent decisions (that is, share with the bad partner and keep with the good partner). (**b**–**d**) Time course of activation (± s.e.m.) for cingulate cortex (Brodmann's area 32; *x, y, z* = –4, 17, 38). No differences between 'share' and 'keep' decisions are observed for (**b**) the neutral partner, but such differences are observed for (**c**) good and (**d**) bad partners, consistent with bias-incongruent behavioral decisions.

conflict evaluation would be recruited more strongly when participants made bias-incongruent choices (that is, keep with the good partner and share with the bad partner) compared with the alternative decision (**Table 3**). As predicted, the contrast of bias-incongruent decisions yielded activation in the cingulate and bilateral insular cortex (**Fig. 5**), regions previously implicated in cognitive control, conflict monitoring and fairness of decisions[29–31]. For the cingulate cortex, we used the extracted mean beta weights in a repeated-measures ANOVA with moral character (good, bad or neutral) and decisions (share and keep) as factors. We found an interaction between moral character and decision ($F_{2,20} = 4.92$, $P < 0.05$), supporting the idea that the cingulate cortex is involved in resolving moral conflicts[32].

DISCUSSION

In the current study, participants played a trust game with three hypothetical partners depicted as being of good, bad or neutral moral character. The perception of moral characteristics biased pre-experimental self-ratings of trust and behavioral choices as participants chose to be more cooperative with the morally good partner. In addition, this bias was successful in modulating neural structures underlying reward and feedback learning. Specifically, activation in the caudate nucleus was observed during the outcome phase for the neutral partners and lottery trials, showing differential responses between positive and negative feedback. These results are consistent with the suggestion that the caudate nucleus is processing feedback information[5,7,8], especially when the feedback is behaviorally relevant[33,34], with the purpose of learning and adapting choices through trial and error[13,14]. The neutral partner results are also consistent with previously observed activation in the caudate nucleus when contrasting reciprocated and unreciprocated cooperation during an economic game in which participants had no prior knowledge about their trading partners[35]. In contrast, the differential response in the caudate nucleus was either not as robust (bad partner) or nonexistent (good partner) when participants had an expectation about the trial on the basis of the partner's perceived moral character. This was especially interesting during good partner interactions, where the number of share choices remained larger than in other conditions, even though frequent violations of expectations occurred.

In a rational choice framework[23], the finding that feedback processing in the human caudate nucleus and subsequent behavior are altered by perceptions of social and moral characteristics might seem puzzling, because individuals are assumed to always be alert to the possibility that exchange partners will behave opportunistically. In Bayesian terms, the character profiles may create a prior belief but feedback should also adjust this prior belief to reflect new evidence. The results suggest that the good profile not only creates a prior belief, but also disrupts the regular encoding of evidence, or learning, from surprising keep outcomes.

Further, the lack of differential responses between positive and negative outcomes when playing with the morally good partner stands in contrast to error-prediction learning during trial-and-error association tasks[24,36]. This learning hypothesis would predict a sharp decrease in the feedback response following violations of expectations, a finding that has been observed in different parts of the human striatum[37,38], more recently in the caudate nucleus[39]. Although the coding of a teaching signal in the caudate nucleus, reflected during feedback processing, was apparent when subjects were playing with the morally neutral partner and, to a lesser extent, with the bad partner, it was not observed during trials with the good partner. Participants instead seemed either less reliant on feedback information or discounted this information. These results indicate that perceptions of moral characteristics can influence choices and the neural mechanisms involved in feedback processing in trial-and-error learning.

During the decision phase, activation was observed in the more ventral portions of the striatum when contrasting risky share decisions with keep choices. This is consistent with the idea that although the dorsal striatum is more involved in processing feedback contingent on choices[13,14], the human ventral striatum has been primarily linked with a role in making predictions[14,26] and anticipating the outcomes of risky decisions[27,28].

In this study, the riskier decisions involved the bad and neutral trials, both of whose mean beta scores suggested a differential response between share and keep. This activation was not merely due to anticipation of a reward, as mean beta scores for share decisions were also higher than those for lottery play in which participants also anticipated a potential gain of the same monetary value. As in the outcome phase, no differential responses were observed for the good partner, perhaps owing to a smaller expectation of risk, further supporting the idea that moral and social perceptions can influence choices. An alternative hypothesis, however, is that participants had a reward reaction to the presentation of the morally good partner, irrespective of decision. This could occur for two potential reasons. First, in previous studies of economics games, activation of the ventral striatum has been reported during cooperation[21] and in response to pictures of previous cooperators[40], suggesting that participants possibly experienced the presentation of the good partner, or the idea of transferring funds to him, as

© 2005 Nature Publishing Group http://www.nature.com/natureneuroscience

© 2005 Nature Publishing Group http://www.nature.com/natureneuroscience

rewarding in itself. A second possibility is that the reward response elicited by the good partner is owing to the participant's experience of possible revenge with a keep decision, because this partner violated trust on half the trials[41]. Although it is unclear what specifically the ventral striatum is responding to (that is, anticipation or prediction of risky decision or reward response owing to presentation of face or opportunity for revenge), reward-related activation in this region may also be modulated by moral and social perceptions.

Our manipulation of the perception of moral characteristics was evident in the behavioral choices, as participants chose to share more with the morally good partner and keep more with the morally bad partner. This bias suggests that cognitive mechanisms involved in conflict processing are recruited when subjects are making a bias-incongruent choice, such as choosing to keep when playing the good partner. Such contrast revealed activation in a number of areas (**Table 3**), including the following: the cingulate cortex, previously linked to conflict monitoring[29], cognitive control[29,31] and resolving difficult moral conflict that may involve personal violations of values[32]; and bilateral insular cortex, implicated in general uncertainty and arousal responses[42] and fairness of decisions[30]. Taken together, these results suggest that although moral information can modulate reward- and feedback-based learning systems in the human brain, the neural circuitry involved in conflict monitoring processes remain unaltered, as participants' analysis of difficult moral decisions are consistent with previous conflict monitoring studies.

In the current experiment, participants' choice behavior in an iterated version of the trust game was influenced by a previous bias regarding moral character. The data suggest that a signal linked to trial-and-error feedback processing, which serves to adapt and optimize choices, was observed in the caudate nucleus during conditions where no information (that is, lottery trials) or little relevant information (that is, neutral partner trials) was available. However, the availability of prior information about moral character diminished the differential signal observed in the caudate. This led participants to discount feedback information and not adapt their choices accordingly, despite showing declarative evidence of learning (as indicated by participant's pre- and post-experiment trustworthiness self-ratings). What remain unanswered are the following questions: (i) what neural mechanisms are linked to this modulation of the teaching signal observed in the caudate nucleus, and (ii) what neural mechanisms are underlying the learning as expressed by changes in the explicit ratings of trustworthiness?

Although the current study cannot provide conclusive answers to these questions, it is clear from both the human and animal literature that there are multiple, interacting systems linked to learning and memory[2,43–46] and that different tasks rely on these systems to varying degrees. In particular, it has been suggested that the caudate feedback learning mechanism may be less involved in tasks where declarative, hippocampal-dependent knowledge can be used to guide behavior[2,44]. This raises the possibility that the lack of observed differential activity in the striatum, as well as the declarative learning shown at the end of the experiment (pre- and post-experiment self-ratings), are hippocampally mediated. We did not observe activation in the hippocampus in our analysis at the thresholds selected for significance. This is perhaps due to the fact that the design used was not optimized to observe hippocampal-dependent learning throughout the task. Additional experiments will be needed to further investigate this potential interaction between the caudate feedback learning system and the hippocampus-declarative memory system during economic interactions.

Moral assessment is a complex domain that needs to be considered when investigating human interactions and decision-making. Future studies are necessary to understand why certain behaviors are affected by a sense of moral obligation. Self-interest alone cannot explain, for example, why people would leave tips in a restaurant they will never visit again. The present study suggests that moral and social perceptions can modulate neural mechanisms associated with feedback and reward processing and cognitive control, thereby influencing our day-to-day choices.

METHODS

Participants and procedure. Fourteen right-handed volunteers participated in this study (8 male, 6 female; average age: mean = 26.64, s.d. = 4.11). One participant was removed from further analysis because of scanner malfunction, and another because of a failure to comprehend instructions (as assessed through post-experimental forms, behavioral results and self-assessment). Data acquired from the 12 remaining participants was included in the analysis. Participants responded to posted advertisements and gave informed consent according to the New York University's Committee of Activities Involving Human Subjects.

Participants were instructed that they were playing a trust game and that they would be playing the three hypothetical partners described in three separate biographies (see **Supplementary Note**). Each partner was introduced with a picture (white, neutral, male faces taken from the NimStim Face Stimulus Set; N. Tottenham, A. Borscheid, K. Ellertsen, D.J. Marcus & C.A. Nelson, *Cogn. Neurosci. Soc. Abstr.* 2002) and a biography describing his characteristics and a recent noteworthy event. Two of the biographies were constructed so as to depict partners with exaggerated moral qualities: one praiseworthy (an English graduate student and volunteer inner-city teacher who rescued a friend from a fire during a crowded concert), and the other suspect (a business graduate student who had been arrested for trying to sell tiles of the space shuttle Columbia on an Internet auction site). The third biography described an engineering graduate student who narrowly missed a doomed flight, but contained no information relevant for assessing his moral character. Participants were instructed that the fictional partners may or may not play according to their described personality or moral characteristics. In fact, each partner had the same 50% reinforcement rate.

Each trial proceeded as follows (**Fig. 1**). The trial was divided into a decision phase and an outcome phase. During the decision phase, participants viewed the name and face of the partner and the options to keep or share, for 3 s followed by a 12-s interval. During the outcome phase, one of three possible outcomes was displayed, indicating the following: (i) the participant chose to keep, (ii) the partner chose to keep or (iii) the partner chose to share. The feedback was presented for 1 s, followed by a 12-s interval. Participants were told they would also play trials involving a lottery game, intermixed with trust trials. In the lottery, participants viewed a dollar sign and chose between 'No' and 'Yes' for a chance to gain $1.50. There was no penalty for losing the lottery. Thus, as expected, 'No' responses were never recorded. There were 96 interleaved trials, divided into 8 runs of 12 trials each. Participants played with each partner 24 times and played 24 lottery trials.

Before the scanning session, participants filled out a seven-point Likert-scale questionnaire, rating the perceived trustworthiness of the three partners. The same questionnaire and additional assessments of learning were administered following the scanning session. After the experimental session was complete, participants were debriefed and paid according to performance (the monetary sum acquired in four out of eight blocks of trials, chosen randomly). Although trial order was predetermined, outcome and feedback were contingent on performance and varied between participants. There were two counterbalanced trial orders, and all three pictures used to represent the partners' faces were counterbalanced across the study. Trials in which a response was not made in time carried a monetary penalty of $1.00 and were excluded from further analysis. Stimulus presentation and behavioral data acquisition were controlled by a Macintosh computer with PsyScope software[47].

Behavioral analysis. During the behavioral session, we analyzed data pertaining to reaction time and choice ('keep' or 'share' decision). We used paired *t*-tests to compare the percentage of time share and keep decisions were made, for each partner individually and also compared decisions between partners. We conducted similar analyses to compare reaction time data. Finally, to determine

how participants adapted their choices as the session progressed, we performed an analysis of early decisions (trials included in the first two fMRI runs, roughly six per partner) and late decisions (last two fMRI runs), using paired *t*-tests.

For the questionnaire data (acquired using a seven-point scale), we conducted a repeated-measures ANOVA using participants as a random factor ($n = 12$), with type of partner (good, bad and neutral) and time of rating (pre- and post-experiment) as within-subject factors. We also used paired *t*-tests to analyze the final post-experiment question regarding the percentage of time participants believed each partner shared with them.

fMRI acquisition and analysis. We used a 3-Tesla Siemens Allegra scanner to collect structural (T1-weighted MPRAGE: 256 × 256 matrix; FOV = 256 mm; 176 1-mm sagittal slices) and functional images (single-shot gradient echo EPI sequence; TR = 2000 ms; TE = 25 ms; FOV = 192 cm; flip angle = 80°; matrix = 64 × 64; slice thickness = 3 mm). Forty contiguous oblique-axial slices (3 × 3 × 3 mm^3 voxels) parallel to the anterior commissure–posterior commissure (AC-PC) line were obtained. fMRI data was analyzed using Brain Voyager software. Preprocessing included motion correction (six-parameter, three-dimensional motion correction), spatial smoothing (4-mm FWHM), voxel-wise linear detrending, high-pass filtering of frequencies (3 cycles per time course) and normalization to Talairach stereotaxic space[48].

We performed random-effects analyses on the functional data for the decision and outcome phase separately. For the outcome phase, we defined a general linear model (GLM) that included eight regressors: two outcomes (positive or negative) for each of four situations (good, bad, neutral or lottery). We conducted a similar GLM during the decision phase that included seven regressors: two decisions (keep or share) for each of three partners (good, bad or neutral), and a lottery play trial. Statistical maps were created using a threshold of $P < 0.001$ with a cluster threshold of 10 voxels[49]. The primary contrasts of interest was the differential response between all positive and negative feedback during the outcome phase, and the differential response between all share and keep choices during the decision phase. A secondary analysis contrasted bias-incongruent (for example, keep with good partner) and bias-congruent choices during the decision phase.

These overall contrasts yielded several regions, including striatum ROIs. We derived statistics from each functional ROI after identifying the peak of activation and surrounding voxels encompassing 10 mm^3. During the outcome phase, a striatum ROI was present in the right hemisphere and contained two separate peaks. The largest peak of activation was observed in the ventral head of the caudate nucleus ($x, y, z = 13, 8, 1$), which was similar in location to other recent reward-learning studies[11,12]. We then used mean beta weights extracted from this ROI in a repeated-measures ANOVA that probed the interaction between moral character (good and bad versus neutral) and outcome (positive versus negative feedback). To explore the source of the interaction, we conducted two additional ANOVAs comparing the good and bad partner separately with the neutral partner. We further investigated these ROIs using mean beta weights for each predictor, which were compared using one-tailed paired *t*-tests. Finally, we plotted time-series data for target ROIs averaged over trials for each individual playing partner and the lottery.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

The authors wish to acknowledge K. Nearing for assistance and useful discussion, J. Pearson and E. Neaville for technical assistance and S. Ravizza for informative discussion and constructive criticism. This work was supported by the US National Institute of Mental Health, the James S. McDonnell Foundation, the Beatrice and Samuel A. Seaver Foundation and Cornell's S.C. Johnson School of Management.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. White, N.M. Mnemonic functions of the basal ganglia. *Curr. Opin. Neurobiol.* **7**, 164–169 (1997).
2. Packard, M.G. & Knowlton, B.J. Learning and memory functions of the basal ganglia. *Annu. Rev. Neurosci.* **25**, 563–593 (2002).
3. Takikawa, Y., Kawagoe, R. & Hikosaka, O. Reward-dependent spatial selectivity of anticipatory activity in monkey caudate neurons. *J. Neurophysiol.* **87**, 508–515 (2002).
4. Shohamy, D. *et al.* Cortico-striatal contributions to feedback-based learning: converging data from neuroimaging and neuropsychology. *Brain* **127**, 851–859 (2004).
5. Poldrack, R.A. *et al.* Interactive memory systems in the human brain. *Nature* **414**, 546–550 (2001).
6. Filoteo, J.V. *et al.* Cortical and subcortical brain regions involved in rule-based category learning. *Neuroreport* **16**, 111–115 (2005).
7. Elliott, R., Sahakian, B.J., Michael, A., Paykel, E.S. & Dolan, R.J. Abnormal neural response to feedback on planning and guessing tasks in patients with unipolar depression. *Psychol. Med.* **28**, 559–571 (1998).
8. Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C. & Fiez, J.A. Tracking the hemodynamic responses to reward and punishment in the striatum. *J. Neurophysiol.* **84**, 3072–3077 (2000).
9. Cromwell, H.C. & Schultz, W. Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *J. Neurophysiol.* **89**, 2823–2838 (2003).
10. Delgado, M.R., Locke, H.M., Stenger, V.A. & Fiez, J.A. Dorsal striatum responses to reward and punishment: effects of valence and magnitude manipulations. *Cogn. Affect. Behav. Neurosci.* **3**, 27–38 (2003).
11. Delgado, M.R., Miller, M.M., Inati, S. & Phelps, E.A. An fMRI study of reward-related probability learning. *Neuroimage* **24**, 862–873 (2005).
12. Haruno, M. *et al.* A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *J. Neurosci.* **24**, 1660–1665 (2004).
13. Tricomi, E.M., Delgado, M.R. & Fiez, J.A. Modulation of caudate activity by action contingency. *Neuron* **41**, 281–292 (2004).
14. O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
15. Barto, A.G. Adaptive critics and the basal ganglia. in *Models of Information Processing in the Basal Ganglia* (eds. Houk, J.C., Davis, J.L. & Beiser, D.G.) 215–232 (MIT Press, Cambridge, Massachusetts, 1995).
16. King-Casas, B. *et al.* Getting to know you: reputation and trust in a two-person economic exchange. *Science* **308**, 78–83 (2005).
17. Frank, R.H., Gilovich, T. & Regan, D. The evolution of one-shot cooperation. *Ethol. Sociobiol.* **14**, 247–256 (1993).
18. Frank, R.H. *What Price the Moral High Ground?* (Princeton Univ. Press, Princeton, New Jersey, 2004).
19. Camerer, C.F. & Weigelt, K. Experimental tests of a sequential equilibrium reputation model. *Econometrica* **56**, 1–36 (1988).
20. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity, and social history. *Games Econ. Behav.* **10**, 122–142 (1995).
21. Rilling, J. *et al.* A neural basis for social cooperation. *Neuron* **35**, 395–405 (2002).
22. Glimcher, P.W. & Rustichini, A. Neuroeconomics: the consilience of brain and decision. *Science* **306**, 447–452 (2004).
23. Coleman, J.S. *Foundations of Social Theory* (Belknap Press, Cambridge, Massachusetts, 1990).
24. Schultz, W. & Dickinson, A. Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* **23**, 473–500 (2000).
25. Delgado, M.R., Stenger, V.A. & Fiez, J.A. Motivation-dependent responses in the human caudate nucleus. *Cereb. Cortex* **14**, 1022–1030 (2004).
26. O'Doherty, J.P. Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Curr. Opin. Neurobiol.* **14**, 769–776 (2004).
27. Knutson, B., Adams, C.M., Fong, G.W. & Hommer, D. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci.* **21**, RC159 (2001).
28. Breiter, H.C., Aharon, I., Kahneman, D., Dale, A. & Shizgal, P. Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* **30**, 619–639 (2001).
29. Carter, C.S. *et al.* Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science* **280**, 747–749 (1998).
30. Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E. & Cohen, J.D. The neural basis of economic decision-making in the ultimatum game. *Science* **300**, 1755–1758 (2003).
31. Ridderinkhof, K.R., Ullsperger, M., Crone, E.A. & Nieuwenhuis, S. The role of the medial frontal cortex in cognitive control. *Science* **306**, 443–447 (2004).
32. Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M. & Cohen, J.D. The neural bases of cognitive conflict and control in moral judgment. *Neuron* **44**, 389–400 (2004).
33. Zink, C.F., Pagnoni, G., Martin-Skurski, M.E., Chappelow, J.C. & Berns, G.S. Human striatal responses to monetary reward depend on saliency. *Neuron* **42**, 509–517 (2004).
34. Elliott, R., Newman, J.L., Longe, O.A. & William Deakin, J.F. Instrumental responding for rewards is associated with enhanced neuronal response in subcortical reward systems. *Neuroimage* **21**, 984–990 (2004).
35. Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E. & Cohen, J.D. Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport* **15**, 2539–2543 (2004).
36. Schultz, W., Dayan, P. & Montague, P.R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
37. McClure, S.M., Berns, G.S. & Montague, P.R. Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38**, 339–346 (2003).
38. O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H. & Dolan, R.J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329–337 (2003).

© 2005 Nature Publishing Group http://www.nature.com/natureneuroscience

39. Davidson, M.C. *et al.* Differential cingulate and caudate activation following unexpected nonrewarding stimuli. *Neuroimage* **23**, 1039–1045 (2004).
40. Singer, T., Kiebel, S.J., Winston, J.S., Dolan, R.J. & Frith, C.D. Brain responses to the acquired moral status of faces. *Neuron* **41**, 653–662 (2004).
41. de Quervain, D.J. *et al.* The neural basis of altruistic punishment. *Science* **305**, 1254–1258 (2004).
42. Critchley, H.D., Mathias, C.J. & Dolan, R.J. Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron* **29**, 537–545 (2001).
43. Gabrieli, J.D. Cognitive neuroscience of human memory. *Annu. Rev. Psychol.* **49**, 87–115 (1998).
44. Squire, L.R., Knowlton, B. & Musen, G. The structure and organization of memory. *Annu. Rev. Psychol.* **44**, 453–495 (1993).
45. Poldrack, R.A. & Packard, M.G. Competition among multiple memory systems: converging evidence from animal and human brain studies. *Neuropsychologia* **41**, 245–251 (2003).
46. Zola-Morgan, S. & Squire, L.R. Neuroanatomy of memory. *Annu. Rev. Neurosci.* **16**, 547–563 (1993).
47. Macwhinney, B., Cohen, J. & Provost, J. The PsyScope experiment-building system. *Spat. Vis.* **11**, 99–101 (1997).
48. Talairach, J. & Tournoux, P. *Co-planar Stereotaxic Atlas of the Human Brain: An Approach to Medical Cerebral Imaging* (Thieme Medical Publishers, New York, 1988).
49. Forman, S.D. *et al.* Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn. Reson. Med.* **33**, 636–647 (1995).

© 2005 Nature Publishing Group http://www.nature.com/natureneuroscience