Cognition 117 (2010) 261–275

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

# Enhanced old–new recognition and source memory for faces of cooperators and defectors in a social-dilemma game

Raoul Bell *, Axel Buchner, Jochen Musch

*Department of Experimental Psychology, Heinrich-Heine-University, Düsseldorf, Germany*

ARTICLE INFO

*Article history:*
Received 26 November 2009
Revised 27 August 2010
Accepted 29 August 2010

*Keywords:*
Reciprocal altruism
Cheater detection
Cooperation
Source monitoring
Context memory
Cheater recognition

ABSTRACT

A popular assumption in evolutionary psychology is that the human mind comprises specialized cognitive modules for social exchange, including a module that serves to enhance memory for faces of cheaters. In the present study, participants played a trust game with computerized opponents, who either defected or cooperated. In a control condition, no interaction took place. In a surprise memory test, old–new recognition for faces and source memory for the associated cooperative or non-cooperative behavior were assessed. A multinomial model was used to measure old–new discrimination, source memory, and guessing biases separately. Inconsistent with the assumption of a memory mechanism that focuses exclusively on cheating, the present study showed enhanced old–new discrimination and source memory for both cooperators and defectors. Rarity of the behavior strategies within the experiment modulated source memory, but only when the differences in base rates were extreme. The findings can be attributed to a mechanism that focuses on exchange-relevant information and flexibly adapts to take into account the relative significance of this information in the encoding context, which may be more beneficial than focusing exclusively on cheaters.



## 1. Introduction

Social cooperation among unrelated individuals is an interesting phenomenon from an evolutionary point of view. At first glance, it seems obvious that individuals benefit from mutual cooperation. However, cooperation is costly for the individual providing support for other group members. Hence, natural selection would work against individuals who *unconditionally* provide benefits to others (Axelrod & Hamilton, 1981; Cosmides, 1989; Cosmides & Tooby, 1992, 2005; Trivers, 1971). In a *single-shot Prisoner's Dilemma Game*, defecting is the dominant strategy, because in each interaction that involves a cooperator and a defector, the defector benefits at the expense of the cooperator. However, things change when the same players interact repeatedly with each other. In this situation, cooperation can be very successful if it is reciprocal, that is, if cooperation is made contingent on the opponent's behavior in previous encounters (Axelrod & Hamilton, 1981; Cosmides, 1989; Trivers, 1971). Reciprocal strategies in social exchange require certain cognitive prerequisites such as the ability to detect cheaters, the ability to recognize different individuals, and the ability to "store information about the history of one's past exchanges with other individuals (in order to know when to cooperate, when to defect, and when to punish defection)" (Cosmides & Tooby, 1992, p. 177).

According to social contract theory (Cosmides, 1989; Cosmides & Tooby, 1992, 2005), social exchange is of such crucial importance for the individual's fitness that specialized cognitive modules have evolved that help us to deal with social-exchange situations. The cheater-detection module proposed by this theory allows the individual to

* Corresponding author. Address: Institut für Experimentelle Psychologie, Heinrich-Heine-Universität, D-40225 Düsseldorf, Germany. Tel.: +49 211 811 5643; fax: +49 211 811 5037.

*E-mail address:* raoul.bell@uni-duesseldorf.de (R. Bell).

0010-0277/$ - see front matter 
doi:10.1016/j.cognition.2010.08.020

quickly and easily draw inferences on whether someone has cheated in prior exchanges or is about to cheat in future interactions. Several researchers (Chiappe et al., 2004; Mealey, Daood, & Krage, 1996; Oda, 1997) have argued that—to save the individual from social exploitation—this cheater-detection module has to be complemented by memory mechanisms that are sensitive to violations of social contract laws and that enable the individual to learn from previous negative experiences with cheaters. There are a number of studies that have examined whether there is a specialized module for remembering faces of cheaters. In most of these studies (Barclay & Lalumière, 2006; Bell & Buchner, 2010, in press-a, in press-b; Buchner, Bell, Mehl, & Musch, 2009; Chiappe et al., 2004; Mealey et al., 1996; Mehl & Buchner, 2008), the moral status of the faces was manipulated using short descriptions in which the stimulus characters were associated with cheating, trustworthy, or irrelevant behavior. In their pioneering study, Mealey et al. observed that for faces associated with low-status professions, old–new discrimination was better for faces of cheaters than for faces of trustworthy persons. Unexpectedly, the pattern was descriptively in the opposite direction for faces associated with high-status professions. Subsequent studies failed to replicate the face recognition effect (Barclay & Lalumière, 2006; Mehl & Buchner, 2008). However, simply recognizing a face of a cheater as familiar cannot help to avoid cheaters in social exchange. Source memory for faces of cheaters, that is, better memory for the cheating context in which a face was encountered, in contrast, can be instrumental in avoiding exploitation and is therefore beneficial to cooperating individuals. Consistent with these assumptions, a series of experiments in our lab (Buchner et al., 2009) showed that source memory for faces of cheaters was enhanced compared to source memory for other types of faces. These findings provide support for a functional perspective on human memory (Klein, Cosmides, Tooby, & Chance, 2002; Nairne, 2005; Nairne & Pandeirada, 2008).

Although most of the aforementioned studies were interpreted with reference to models of direct reciprocity (Axelrod & Hamilton, 1981; Trivers, 1971), the moral status of the stimulus characters was manipulated by letting participants read about interactions involving third parties. Cheating and trustworthiness had no negative or positive consequences for participants whatsoever, simply because participants were not directly involved in social exchange with the characters. We henceforth refer to this paradigm as the *description paradigm* (Hammerl, 2000). It is unclear whether the pattern of results obtained in the *description paradigm* may generalize to a situation in which participants are directly involved in social exchange with the stimulus characters. Therefore, it is interesting to examine old–new discrimination and source memory for faces of cheaters in an *involvement paradigm*, that is, using a task in which cheating and trustworthy behavior of the stimulus characters has direct negative or positive consequences for participants. The simplest way to do this is to use a *social-dilemma game*.

To date, memory for faces of cheaters was investigated in three studies using a *social-dilemma game* (Barclay, 2008; Oda, 1997; Singer, Kiebel, Winston, Dolan, & Frith, 2004). Two of these studies manipulated social-exchange status by providing participants with third-party reputational information about the stimulus persons. Oda (1997) required participants to imagine that they were one of two prisoners in the original version of the *Prisoner's Dilemma* and provided information about the opponents' strategies in this game (confessing or keeping silent). The opponents' behavior had only imaginary, but no real negative or positive consequences for the participants. Barclay (2008) informed participants that they would have to play a trust game with computerized players. Before the memory test, participants were given photographs of their prospective opponents and were explicitly informed about who will defect and who will cooperate in the game phase. Given that participants knew that they would benefit directly from encoding and rehearsing the strategies associated with the opponents' faces, the instructions used by Barclay can be equated with explicit learning instructions. Therefore, it is possible that conscious, strategic rehearsal may have overridden the evolved learning biases that may affect memory primarily in situations in which learning is less strategic. This suspicion is substantiated by a study of D'Argembeau and Van der Linden (2004), showing that source memory was enhanced for emotional words when compared to neutral words, but only when learning was incidental and participants had no intention to encode the context of the words anyway. In summary, although the studies of Oda and Barclay manipulated social status using social-dilemma games, participants were provided with third-party reputational information rather than with first-party experience. Consequently, their results are inconclusive regarding the prediction derived from models of direct reciprocity (Axelrod & Hamilton, 1981; Trivers, 1971) that participants *spontaneously* encode and remember their opponents' faces and the associated behavior strategies when they are *directly* involved in social exchange with them.

There is only one study we know of that examined memory for cheaters using an *involvement* paradigm. In this fMRI study (Singer et al., 2004), participants played an iterated trust game with computerized players. Participants could trust their opponents by sending them money. The opponents either reciprocated by sending the money back (which resulted in a moderate profit for both players involved) or defected by keeping the money (whereby they maximized their own profit at the expense of the participants). In a control condition, participants saw the faces of the opponents, but no transaction took place. Each participant interacted repeatedly with the same set of opponents that consisted of five cooperators, three defectors, and three irrelevant control faces. The neuroimaging results revealed enhanced activation in brain regions that are commonly associated with emotional processing and social cognition in response to the cooperator and defector faces in comparison with the irrelevant faces. The study also comprised a source memory test. In this test, participants were asked whether a face belonged to a defector, a cooperator or to an irrelevant control person. The raw number of correct source classifications was more accurate for faces of cooperators than for other faces. The main problem with interpreting ad hoc source memory measures such as the raw number of correct source classification is that these measures do not allow one to separately assess old–new discrimination, source memory, and guessing biases. The failure to take guessing into account is especially problematic given that the different proportions of cooperators and defectors in the game phase may have induced a bias towards guessing that a face belonged to a cooperator. Such a guessing bias would selectively increase the number of correct source classifications for cooperator faces. This example shows that it is important to use measurement tools that allow for the independent assessment of old–new discrimination, source memory, and guessing biases. Fortunately, such measurement tools exist in terms of multinomial models of source memory (Batchelder & Riefer, 1990; Bayen, Murnane, & Erdfelder, 1996).

The purpose of Experiment 1 was to examine old–new recognition and source memory for faces of cooperators and defectors using an involvement paradigm in which participants interacted directly with their opponents, and using a multinomial model of source memory (Bayen et al., 1996) that provides an unconfounded assessment of old–new discrimination, source memory and guessing biases. In Experiments 2 and 3, we extend the findings of Experiment 1 by manipulating the base rates of cooperators and defectors within the experiment to see whether we could replicate the finding of Barclay (2008) that the frequency of the opponent types determines memory for faces. We were especially interested in the question whether rarity would enhance memory for information regardless of its social relevance. If there were memory specializations for social exchange, we would expect to find that memory for exchange-relevant information is always better than memory for irrelevant information regardless of the rarity of this information.

## 2. Experiment 1

### 2.1. Method

#### 2.1.1. Participants

Participants were 89 persons (58 women), most of whom were students at Heinrich-Heine-University of Düsseldorf. Their age ranged from 19 to 63 years ($M = 26$). All participants were tested individually.

#### 2.1.2. Materials and procedure

Seventy-two facial photographs of males (256 bit, $116 \times 164$ pixel grayscales) were randomly assigned to two sets of 36 photographs each (henceforth Sets 1 and 2). Thirty-six of these photographs (Set 1 or 2, counterbalanced across participants) were presented in random order during the trust game.

In the game phase, each participant played a trust game with 36 opponents (12 defectors, 12 cooperators, and 12 irrelevant persons). The participant was represented by a contour of either a man or a woman—matching the participant's sex—that was presented on the left side of the screen (Fig. 1). On the right side of the screen, opposite to the participant's contour, a facial photograph of the opponent was shown. The current account balance of the participant and that of the opponent were shown at the top left and right side of the screen, respectively.

Each participant started with a deposit of 450 cents. On the participant's side, there was a computer button which read "to be invested". Upon a click on this button, a box appeared from which the participant could choose how much to invest. The participant could invest 15, 30, 45, or 60 cents. Once confirmed, the selected amount was shown in an arrow, the size of which was adjusted to the amount of the investment. The arrow remained on the left side of the screen for 1 s and moved towards the center of the screen within 1.5 s, where it remained for the rest of the trial. One second after it had stopped, the opponent's investment appeared in an arrow on the right side of the screen. A cooperator invested the same amount of money as the participant. A defector invested much less than the participant (1/3 of the amount the participant had invested). If the opponent was an irrelevant person, the arrow was empty. Again, the arrow was presented for 1 s before it moved towards the center of the screen, where it remained for the rest of the trial.

The sum of investments, the profit, and the total sum appeared one after another in 1 s intervals at the center of the screen. The profit was always 20% of the investments. The sum of the investments and the profit were added up and the total sum was split up between the participant and the opponent. Both received half of the total sum regardless of their investments. The corresponding amounts were presented in two arrows in the center of the screen, one on the participant's half and the other on the opponent's half. If the opponent was irrelevant, the participant got back what he or she had invested and the opponent's arrow was again empty. One second after the total sum had appeared, the arrow containing the opponent's share moved towards the opponent's photograph. One second later, the arrow containing the participant's share moved towards the participant's contour. On both sides, the amount of gain or loss (i.e., the difference between each opponent's investment and each opponent's share of the total sum) appeared simultaneously underneath the pictures in green or red font color, respectively. Finally, the updated account balance appeared in black font color. At the end of each trial, a short summary of the interaction was shown at the bottom of the screen. The participant initiated the next trial by clicking on a "continue" button.

Thus, if a participant would opt to invest 30 cents, a cooperative opponent would also invest 30 cents. Accordingly the sum of investments would be 60 cents. The profit would be 12 cents, and the total return would be 72 cents. Thus, each opponent would receive 36 cents as an equal share of the total sum of return. Both the participant and the opponent would gain 6 cents. A defecting opponent, in contrast, would only invest 10 cents. Accordingly, the total sum of investment would be limited to 40 cents, and the profit and the total sum of return would amount to 8 cents and 48 cents, respectively. The participant and the defector would receive 24 cents as their share of the total sum. Thus, the participant would incur a loss of 6 cents, whereas the defector would gain 14 cents. Note that the participant's gain in the cooperator condition is thus as

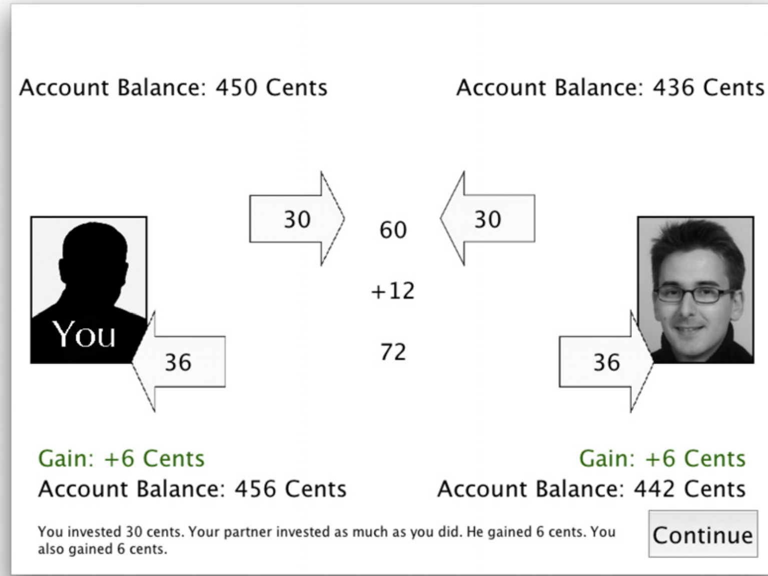


**Fig. 1.** Screen shot of the trust game. On the left side of the screen, the participant's contour is shown. On the right side on the screen, a facial photograph of the (cooperative) opponent is shown. Above the participant's contour and the opponent's face, the account balance of the participant and the opponent is shown respectively. The numbers inside the arrows refer to the participant's and the opponent's investments and the participant's and the opponent's share, respectively. In the center of the screen, the sum of the investments, the profit, and the total sum are presented. Below the pictures, the gain and the updated account balance are shown. At the bottom of the screen, a short verbal summary of the interaction is presented.

large as the participant's loss in the defector condition. In the irrelevant control condition, participants could not gain nor lose any money. Participants knew that they played for real money, and that they would be paid out the amount scored at the end of the experiment.

The encoding phase was directly followed by the test phase. Here, participants saw a random sequence of 72 faces, half of which had been presented in the trust game (Set 1 or 2, depending on the trust game assignment), and half were new (Set 2 or 1). The 36 "old" faces belonged to the 12 defectors, 12 cooperators, and 12 irrelevant persons presented in the trust game. Each trial started with a headline ("How likable do you find this person?") and a photograph. The likability rating scale appeared 1.5 s later. After the rating a new headline appeared ("Is this face old or new?"), followed by an "old" and a "new" checkbox, one of which participants selected to indicate that they had seen a face during the trust game or not. Following an "old" judgment and a click on the *continue* button, a headline appeared that read "Was this person..." together with three checkboxes labeled "...a cheater, because he invested less money than you, so that for you it was a loss, and for him a good profit", "...a trustworthy person, because he invested as much money as you did, so that both of you made a profit", and "neither a cheater nor a trustworthy person, because the transaction did not take place". Participants were required to use these checkboxes to judge the behavior of the opponents in the trust game. Selecting one of these checkboxes and then clicking the *continue* button initiated the next trial.

### 2.1.3. Design

The within-subject independent variable was opponent type (*defector*, *cooperator*, *irrelevant*). The dependent measures were likability ratings, old–new discrimination in terms of the sensitivity measure of the two-high threshold model of signal detection, $P_r$ (Snodgrass & Corwin, 1988), and source judgments given an "old" judgment. In the experiments reported in this article, a multivariate approach was used for all general linear model within-subjects comparisons. In our application, all multivariate test criteria correspond to the same (exact) $F$ statistic, which is reported. The level of $\alpha$ was set to .05 for all analyses.

Given a sample size of $N = 89$, and $\alpha = .05$, the power to detect a difference between the source memory parameters for defectors and cooperators with an effect size of $w = 0.05$—which is in the order of magnitude of the source memory effect observed by Buchner et al. (2009)—was .95. All power calculations were conducted using G•Power (Faul, Erdfelder, Lang, & Buchner, 2007).

## 2.2. Results

### 2.2.1. Trust game investments

On average, participants invested 30 cents. Given that participants could choose to invest 15, 30, 45, or 60 cents, this finding suggests that, on average, participants pursued a cautious (low risk, low gain) investing strategy. There was no effect of opponent type on the amount invested, which was to be expected given that participants did not know whether an opponent would defect or cooperate when they decided how much to invest.

### 2.2.2. Old–new discrimination

Table 1 shows old–new discrimination in terms of $P_r$ (the sensitivity measure of the two-high threshold model), which is calculated by subtracting the false alarm rate from the hit rate. We report $P_r$ as a sensitivity measure because it was favorably evaluated in validation studies (Snodgrass & Corwin, 1988) and avoids the problem of undefined values that comes with using $d'$. Old–new discrimination differed significantly as a function of opponent type, $F(2, 87) = 15.25$, $p < .001$, $\eta^2 = .26$. Orthogonal contrasts showed that old–new discrimination did not differ between defector and cooperator faces, $F(1, 88) = 2.66$, $p = .11$, $\eta^2 = .03$, and was worse for irrelevant faces than for faces of defectors and cooperators, $F(1, 88) = 28.07$, $p < .001$, $\eta^2 = .24$.

### 2.2.3. Test-phase likability ratings

Test-phase likability ratings (Fig. 2) differed as a function of opponent type, $F(2, 87) = 53.74$, $p < .001$, $\eta^2 = .55$. Post hoc contrasts showed that defector faces were less likable than irrelevant faces, $F(1, 88) = 7.16$, $p < .01$, $\eta^2 = .08$, and that irrelevant faces were less likable than cooperator faces, $F(1, 88) = 68.42$, $p < .001$, $\eta^2 = .44$.

### 2.2.4. Source memory

As discussed in the introduction, there are different ways to measure source memory. We start by reporting the conditional source identification measure (CSIM), that is, the number of correct source classifications conditionalized on the number of faces given an "old" classification. The CSIM for a defector face is given, for instance, by $f_{DefectorDefector}/(f_{DefectorDefector} + f_{DefectorCooperator} + f_{DefectorIrrelevant})$, where $f_{ij}$ is the frequency of responses of type $j$ to items of type $i$. Ad

**Table 1**
Old–new discrimination in terms of $P_r$ for Experiments 1–3 and test-phase likability ratings and conditionalized source identification for Experiment 3.

| | Opponent type | | |
|---|---|---|---|
| | Defector | Irrelevant | Cooperator |
| *Old–new discrimination* | | | |
| *Experiment 1* | | | |
| | 0.62 (*SE* = 0.03) | 0.55 (*SE* = 0.03) | 0.65 (*SE* = 0.02) |
| *Experiment 2* | | | |
| 60% Defectors group | 0.52 (*SE* = 0.03) | 0.52 (*SE* = 0.03) | 0.56 (*SE* = 0.03) |
| 60% Cooperators group | 0.58 (*SE* = 0.03) | 0.55 (*SE* = 0.03) | 0.55 (*SE* = 0.03) |
| *Experiment 3* | | | |
| 80% Defectors group | 0.60 (*SE* = 0.04) | | 0.66 (*SE* = 0.04) |
| 80% Cooperators group | 0.62 (*SE* = 0.04) | | 0.61 (*SE* = 0.03) |
| *Test-phase likability ratings (Experiment 3)* | | | |
| 80% Defectors group | 2.60 (*SE* = 0.09) | | 2.95 (*SE* = 0.09) |
| 80% Cooperators group | 2.76 (*SE* = 0.09) | | 3.02 (*SE* = 0.09) |
| *Conditionalized source identification measure (Experiment 3)* | | | |
| 80% Defectors group | 0.87 (*SE* = 0.03) | | 0.50 (*SE* = 0.03) |
| 80% Cooperators group | 0.54 (*SE* = 0.03) | | 0.78 (*SE* = 0.03) |

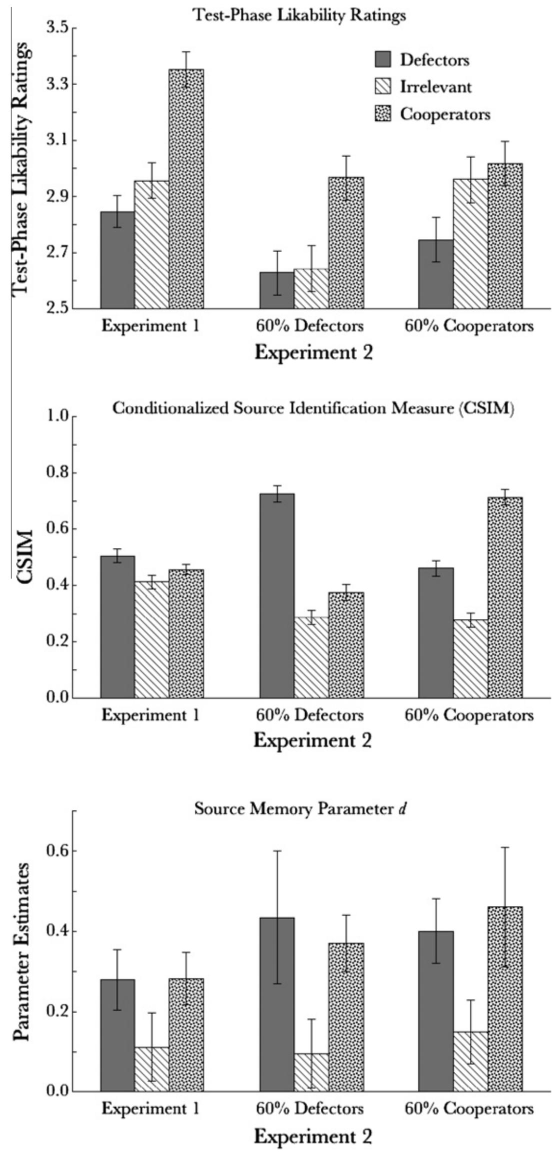


**Fig. 2.** Test-phase likability ratings and memory measures as a function of opponent type (defectors, irrelevant, cooperators). The results for the 60% defectors and the 60% cooperators group of Experiment 2 are displayed separately. *Test-phase likability ratings*: Test-phase likability ratings on a scale from 1(not likable at all) to 6 (extremely likable). Error bars represent the standard errors of the means. *Conditionalized Source Identification Measure (CSIM)*: correct source classifications conditionalized on the number of "old" judgments. Error bars represent the standard errors of the means. *Source memory parameter d*: parameter estimates for the source memory parameter for faces of defectors ($d_{Defect}$), for irrelevant faces ($d_{Irrelevant}$), and for faces of cooperators ($d_{Coop}$). The parameters represent conditional probabilities of correct source identification given correct old–new discrimination. Error bars represent the .95 confidence intervals. Note that different sets of faces had been used in Experiments 1 and 2.

hoc measures such as the CSIM, however, are known to be problematic because they may confound source memory with guessing biases (Bayen et al., 1996; Bröder & Meiser, 2007). A multinomial analysis of the source memory data is to be preferred because it allows us to assess old–new discrimination, source memory, and guessing biases separately. We nevertheless report the CSIM for two reasons. First, it has often been used in previous source memory studies. Second, some inconsistencies in the literature can be attributed to the failure to use source memory measures that take guessing into account. Demonstrating how ad hoc measures such as the CSIM are influenced by source memory *and* guessing biases may help to explain these inconsistencies.

Fig. 2 displays the CSIM for all three opponent types separately. The analysis of the CSIM revealed a significant main effect of opponent type, $F(2, 87) = 4.29$, $p = .02$, $\eta^2 = .09$. Orthogonal contrasts showed that the CSIM did not differ between defector and cooperator faces, $F(1, 88) = 2.94$, $p = .09$, $\eta^2 = .03$. However, the CSIM was higher for defector and cooperator faces than for irrelevant faces, $F(1, 88) = 5.83$, $p = .02$, $\eta^2 = .06$.

These results were confirmed by the multinomial analysis. The source memory model used in Experiment 1 is depicted in Fig. 3. The model contains twelve free parameters, each of which represents the probability with which certain cognitive processes occur. To illustrate, parameter $D_{Defect}$ represents the probability of recognizing a defector face as old. Parameter $d_{Defect}$ represents the conditional probability of also remembering that the face belonged to a defector. If the source of a recognized face is not known (with probability $1 - d_{Defect}$), it may be guessed that the face belonged to a defector with probability $a_{DefectCoop} \cdot a_{Defect}$, to a cooperator with probability $a_{DefectCoop} \cdot (1 - a_{Defect})$, or to an irrelevant person with probability $(1 - a_{DefectCoop})$. If a defector face is not recognized as old (with probability $1 - D_{Defect}$), it may still be guessed, with probability $b$, that the face is old. For these faces, it may be guessed that the face belongs to a defector with probability $g_{DefectCoop} \cdot g_{Defect}$, to a cooperator with probability $g_{DefectCoop} \cdot (1 - g_{Defect})$, or to an irrelevant person with probability $(1 - g_{DefectCoop})$. If a defector face is neither recognized as old, nor guessed to be old (with probability $1 - b$), it is incorrectly judged to be new. Analogous statements hold for the model trees for cooperator faces, irrelevant faces, and new faces. Based on these model equations and the empirically observed sample responses to the different types of faces, it is possible to estimate the model parameters using standard computer programs (Moshagen, 2010).

To obtain an identifiable base model, we started the analysis by setting the parameter representing the proba-

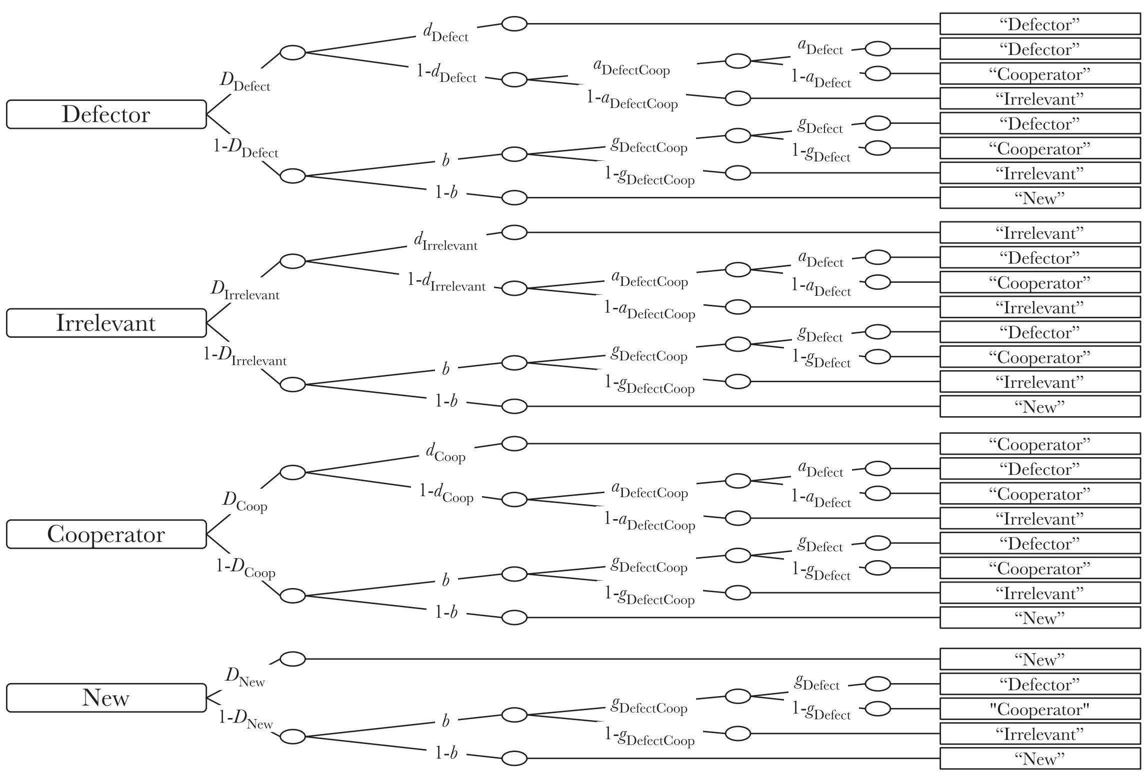


**Fig. 3.** Bayen et al.'s (1996) source memory model as adapted for Experiments 1 and 2. Rounded rectangles on the left side represent the types of faces presented (*defectors*, *irrelevant characters*, *cooperators*, *new faces*). Letters along the links represent the probabilities with which certain cognitive states occur ($D$: probability of correctly identifying a face as old or new; $d$: source memory in the sense of remembering the context of encountering a face that was detected as old; $b$: probability of guessing that a non-recognized face is old; $a$: probability of guessing that a recognized face was encountered in a particular context; $g$: probability of guessing that a non-recognized face that was guessed to be old was encountered in a particular context). Rectangles on the right side represent the categories of participant's judgments.

bility of detecting new faces as new to be equal to the face recognition parameter for irrelevant faces (i.e., $D_{Irrelevant} = D_{New}$). The assumption that the probability for recognizing new faces as "new" is equal to the probability of recognizing old faces as "old" is the standard assumption of the two-high threshold model of signal detection and is empirically justified by the mirror effect (Glanzer, Adams, Iverson, & Kim, 1993). Validation studies have shown that two-high threshold models that make this assumption are superior to one high threshold models (Bayen et al., 1996; Snodgrass & Corwin, 1988). The base model that incorporates this assumption (henceforth Base Model 1) fit the data extremely well, $G^2(1) = 0.56$, $p = .46$.

First, we tested whether old–new discrimination for faces of defectors differs from old–new discrimination for faces of cooperators by imposing, on Base Model 1, the restriction that $D_{Defect} = D_{Coop}$. The *increase* in model misfit as a result of the additional restriction is expressed in the goodness-of-fit statistic $\Delta G^2$, which is asymptotically $\chi^2$ distributed with one degree of freedom. The restriction was compatible with the data, $\Delta G^2(1) = 1.87$, $p = .17$, forcing us to conclude that old–new discrimination for defector and cooperator faces did not differ. The set of restrictions applied so far were combined into Base Model 2, which also fit the data well, $G^2(2) = 2.43$, $p = .30$. Next, we tested the hypothesis that old–new discrimination is better for faces of cooperators and defectors than for irrelevant faces by imposing the restriction that $[D_{Defect} = D_{Coop}] = D_{Irrelevant}$ on Base Model 2, which was clearly incompatible with the data, $\Delta G^2(1) = 24.09$, $p < .001$. Thus, old–new discrimination was better for defector faces and cooperator faces than for irrelevant faces ($[D_{Defect} = D_{Coop}] = .66$ [$CI$ = 0.63–0.69]; $[D_{Irrelevant} = D_{New}] = .55$ [$CI$ = 0.52–0.58]). This mirrors the analysis of the old–new discrimination scores reported above.

Fig. 2 displays the source memory parameter estimates. The null hypothesis that there is no difference between source memory for defectors and cooperators can be implemented by imposing, on Base Model 2, the restriction that $d_{Defect} = d_{Coop}$, which was clearly compatible with the data, $\Delta G^2(1) < 0.01$, $p = .95$. Source memory for cooperators was as good as source memory for defectors. The restrictions applied so far were combined into Base Model 3, which also fit the data well, $G^2(3) = 2.43$, $p = .49$. Next, we tested whether source memory for irrelevant faces was worse than source memory for faces of cooperators and faces of defectors, by imposing, on Base Model 3, the restriction that $[d_{Defect} = d_{Coop}] = d_{Irrelevant}$, which was clearly incompatible with the data, $\Delta G^2(1) = 10.81$, $p < .001$. Source memory for irrelevant faces was worse than source memory for faces of cooperators and defectors.

### 2.3. Discussion

Experiment 1 revealed a clear memory advantage for exchange-relevant information in comparison to irrelevant information. Participants were better at recognizing faces of defectors and cooperators than irrelevant control faces and also better at remembering the contexts in which these faces were presented. These results fit to other results showing that socially relevant information is especially well remembered (e.g., O'Gorman, Wilson, & Miller, 2008). Given that this information may be instrumental in deciding whether to engage in social exchange with these individuals or not, the results support a functional view of human memory (Klein et al., 2002; Nairne, 2005; Nairne & Pandeirada, 2008).

An interesting aspect of the results of Experiment 1 is that they differ from the results obtained in the *description* paradigm. With respect to old–new face recognition, studies using third-party reputational information to manipulate social-exchange status yielded mixed results (e.g., Barclay, 2008; Bell & Buchner, in press-a; Buchner et al., 2009; Mealey et al., 1996; Mehl & Buchner, 2008; Oda, 1997), but the prevalent finding is that old–new face discrimination is not affected by third-party reputational information at all (Barclay & Lalumière, 2006; Bell & Buchner, in press-a; Buchner et al., 2009; Mehl & Buchner, 2008). In contrast, the results of Experiment 1 show that faces that were associated with exchange-relevant information were better recognized than irrelevant control faces. This finding may be explained by assuming that information that stems from first-party experience is more emotionally involving, and therefore more likely to lead to better encoding of the stimulus face, than third-party reputational information.

With respect to source memory, several studies using the description paradigm yielded evidence for a memory advantage for faces of cheaters (Bell & Buchner, in press-a, in press-b; Buchner et al., 2009). A rather surprising finding is that this tendency to remember the negative behavior descriptions was replaced by an advantage for both negative (defectors) and positive (cooperators) encounters in Experiment 1. This finding seems to suggest that the relative importance of cooperation and cheating changes depending on whether people are directly involved with the stimulus characters or not. When participants experience cooperation directly, cooperation may be just as important to remember as cheating. To anticipate, this data pattern was also obtained in Experiments 2 and 3, which is why we discuss this aspect of the results more thoroughly in Section 5.

Even though this pattern of results is inconsistent with hypotheses that were previously derived from social contract theory (e.g., Oda, 1997), it is obvious that good old–new recognition and source memory for faces of cooperators can be beneficial for cooperating individuals. Models of direct reciprocity (Axelrod & Hamilton, 1981; Trivers, 1971) predict that cooperating individuals should learn to adjust their behavior according to their opponent's behavior in previous social encounters. To know when to cooperate and when to defect on basis of previous learning experiences, one has to remember both the cooperators and the cheaters (see also Cosmides & Tooby, 1992).

Experiments 2 and 3 were motivated by the assumption that there might be situations in which it is more beneficial to have good memory for cooperators, and other situations in which it is more beneficial to have good memory for cheaters. Specifically, the ratio of cooperators to defectors in a population may influence the relative importance of having good memory for cheaters versus cooperators (Aktipis, 2006; Barclay, 2008). When cooperation is the rule, it may be more beneficial to have a general tendency to engage in social exchange, and to remember the few individuals that violate the social norm. When cooperation is rare, it may be more beneficial to restrict cooperation to those individuals that have proven to be trustworthy in previous encounters. This requires good memory for cooperators (Aktipis, 2006; Barclay, 2008). Barclay (2008) examined this question directly by manipulating the ratio of cooperators to defectors in the trust game. Consistent with the results of Experiment 1, he found that old–new discrimination and source memory for faces of defectors and cooperators was equally good when participants had encountered 50% defectors and 50% cooperators. When the proportion of defectors and cooperators differed, old–new discrimination and source memory were better for faces associated with the rare behavior. Barclay suggested that his results could be explained by a mechanism that tracks socially relevant information and that is sensitive to the relative frequency of this information.

There are at least three reasons why it may be important to replicate Barclay's (2008) finding using the present *involvement* paradigm. First, Barclay examined how well participants remembered third-party reputational information about how their opponents would behave if they were to interact with them in the future. As already pointed out in the *Introduction*, these instructions can be equated with explicit instructions to remember who will cooperate and who will defect later in the experiment. Consequently, the rarity effects observed by Barclay could be interpreted as the consequence of conscious learning strategies that may be only applied when learning is enforced by the instructions. If this were the case, then these effects should disappear when learning is incidental and thus less strategic. To show that these effects can indeed be ascribed to a frequency-sensitive memory mechanism that is specialized to exchange-relevant information, as proposed by Barclay, it is necessary to replicate the findings using an *involvement* paradigm, in which participants directly interact with the stimulus persons and have no explicit learning instructions.

Second, Barclay's (2008) study lacks a critical control condition in which the faces are *not* associated with socially relevant information. Although Barclay is sympathetic to the idea that information that is relevant to social exchange is especially well remembered, it is impossible to show this with his experimental design. As a consequence, the results of this study could be ascribed to a domain-general mechanism that enhances memory for unusual or rare information regardless of its content. Indeed, a number of results show that rare or unexpected information stands out and is therefore well attended (e.g., Parmentier, Elford, Escera, Andres, & San Miguel, 2008; Schröger & Wolff, 1998) and remembered (Hunt & Worthen, 2006; Schmidt, 1991). To examine whether there is domain-specificity in social exchange, it is necessary to include a control condition to show that exchange-relevant information is indeed better remembered than other information.

Lastly, Buchner et al. (2009) obtained findings in the description paradigm that seem to be inconsistent with the rarity effect observed by Barclay (2008). In three of the four experiments reported by Buchner et al., the exceptionality of the descriptions of cheating, trustworthy, and irrelevant behavior was manipulated. In all experiments, source memory for cheating was enhanced, whereas exceptionality had no effect on memory at all. These findings led to the conclusion that social relevance is much more important in determining memory than exceptionality, and that the source memory advantage for faces of cheaters cannot be attributed to domain-general mechanisms that focus on unusual or exceptional information.

In Experiment 2, we examined this theoretically important question by manipulating the base rates of the behavioral strategies (defection, cooperation) in the trust game. We were especially interested in finding out whether a tendency towards remembering rare stimuli would be found even for socially neutral (irrelevant) faces if this information was rare or whether the social relevance of the faces would turn out to be more important in determining old–new discrimination and source memory.

## 3. Experiment 2

### 3.1. Method

#### 3.1.1. Participants

Participants were 113 persons (74 women). Their age ranged from 19 to 68 years ($M = 25$). The participants were randomly assigned to two groups (see below).

#### 3.1.2. Materials and procedure

Materials and procedure were identical to Experiment 1 with the following exceptions. A total of 96 facial photographs of males were used that were taken from the Psychological Image Collection at Stirling (PICS; http://pics.psych.stir.ac.uk/).

In the trust game, 60 randomly selected pictures were used. In the 60% defectors group, 36 opponents (60%) defected, 12 opponents (20%) cooperated, and there were 12 irrelevant faces (20%). In the 60% cooperators group, 36 opponents (60%) cooperated, 12 opponents (20%) defected, and there were 12 irrelevant faces (20%). The trust game was split up in four consecutive blocks of 15 trials that consisted of 9 trials of the frequent condition, and 3 trials of each of the other two conditions. The order of trials within a block was randomly determined.

In the test phase, 72 facial photographs were presented. Thirty-six of these photographs were new, and 36 were old. Of these, 12 had been associated with defectors, 12 had been associated with cooperators, and 12 had been associated with irrelevant faces in the trust game. The 12 faces of the frequent condition were selected by randomly drawing 3 of these faces from each of the four blocks.

#### 3.1.3. Design

The between-subjects independent variable was the proportion of defectors and cooperators in the trust game (*60% defectors*, *60% cooperators*) and the within-subject variable was opponent type (*defector*, *cooperator*, *irrelevant*). The dependent variables were test-phase likability ratings, old–new discrimination, and source judgments given an "old" judgment. Given a sample size of $N = 113$, $\alpha = .05$, effects of size $w = 0.04$ could be detected for the source memory effect with a probability of $1 - \beta = .95$.

## 3.2. Results

### 3.2.1. Trust game investments

A 2 × 3 MANOVA with group and opponent type as independent variables showed that participants in the 60% defectors group invested less (24 cents) than the participants in the 60% cooperators group (32 cents), $F(1, 111) = 17.06$, $p < .001$, $\eta^2 = .13$. As was to be expected, there was no significant main effect of opponent type and no interaction between the two variables. It may be interesting to compare the investment levels across experiments to see whether high proportions of defectors and cooperators affected investment levels significantly. Participants in the 60% defectors group invested significantly less than participants in Experiment 1, $t(144) = 3.51$, $p < .01$. The investment level in the 60% cooperators group increased only descriptively in comparison to Experiment 1, $t(143) = 1.27$, $p = .21$. This finding suggests that with respect to their willingness to invest into the trust game, participants were more risk averse when defectors were common than they were risk seeking when cooperators were common. This fits to the finding of Experiment 1 that participants pursued a cautious strategy when deciding how much to invest.

### 3.2.2. Old–new discrimination

Old–new discrimination (Table 1) did not differ as a function of group, $F(1, 111) = 0.44$, $p = .51$, $\eta^2 < .01$, or opponent type, $F(2, 110) = 0.52$, $p = .60$, $\eta^2 < .01$. There was also no significant interaction between the two variables, $F(2, 110) = 2.79$, $p = .07$, $\eta^2 = .05$. Descriptively, defectors were slightly better recognized than cooperators in the 60% cooperators group, but cooperators were slightly better recognized than defectors in the 60% defectors group.

### 3.2.3. Test-phase likability ratings

Test-phase likability ratings (Fig. 2) did not differ as a function of group, $F(1, 111) = 2.49$, $p = .12$, $\eta^2 = .02$. However, there was a significant effect of opponent type, $F(2, 110) = 25.64$, $p < .001$, $\eta^2 = .32$, that was qualified by a group × opponent type interaction, $F(2, 110) = 5.17$, $p < .01$, $\eta^2 = .09$. Orthogonal contrasts showed that in the 60% defectors group likability ratings were higher for the cooperator faces than for other types of faces, $F(1, 56) = 33.83$, $p < .001$, $\eta^2 = .38$, but did not differ between defector faces and irrelevant faces, $F(1, 56) = 0.07$, $p = .79$, $\eta^2 < .01$. In the 60% cooperators group likability ratings were lower for the defector faces than for other types of faces, $F(1, 55) = 28.41$, $p < .001$, $\eta^2 = .34$, but did not differ between the cooperator faces and the irrelevant faces, $F(1, 55) = 0.77$, $p = .38$, $\eta^2 = .01$.

### 3.2.4. Source memory

The CSIM (Fig. 2) did not differ as a function of group, $F(1, 111) = 0.69$, $p = .41$, $\eta^2 < .01$. However, there was a significant effect of opponent type, $F(2, 110) = 97.80$, $p < .001$, $\eta^2 = .64$, that was qualified by a significant group × opponent type interaction, $F(2, 110) = 68.20$, $p < .001$, $\eta^2 = .55$. The interaction was due to the fact that the CSIM was largest for defectors in the 60% defectors group and largest for cooperators in the 60% cooperators group.

The situation so far is rather interesting. On the one hand, the analysis of the CSIM seems to suggest that source memory is *best* for the most common behavior (i.e. for defectors in the 60% defectors group and for cooperators in the 60% cooperators group). On the other hand, the test-phase likability ratings do not differ between defector faces and irrelevant faces in the 60% defectors group, and do not differ between cooperator faces and irrelevant faces in the 60% cooperators group (see Fig. 2). How can this puzzle be solved?

Both the CSIM and the test-phase likability ratings may be influenced by several different cognitive processes such as old–new discrimination, source memory, and guessing biases. It is therefore important to analyze the source memory data using the multinomial source memory model which ensures an unconfounded measurement of these components. For an analysis of the present data we need two sets of the four model trees depicted in Fig. 3, one set for the 60% defectors group, and another set for the 60% cooperators group. Correspondingly, there are now also two sets of the parameters that occur in these trees. For instance, there is one parameter representing source memory for defectors in the 60% defectors group, $d_{Defect|60\%Defect}$, and one parameter representing source memory for defectors in the 60% cooperators group, $d_{Defect|60\%Coop}$.

Just as in Experiment 1, we set the parameters representing the probability of detecting new faces as new to be equal to the face recognition parameters for irrelevant faces (i.e., $D_{Irrelevant|60\%Defect} = D_{New|60\%Defect}$, and $D_{Irrelevant|60\%Coop} = D_{New|60\%Coop}$) to obtain an identifiable base model. To simplify the analysis of the guessing parameters, we also assumed that the probability of guessing that a face belonged to either a defector or a cooperator and the probability of guessing that a face belonged to a defector are independent of whether the face was recognized as "old" or not (i.e., $a_{DefectCoop|60\%Defect} = g_{DefectCoop|60\%Defect}$, $a_{DefectCoop|60\%Coop} = g_{DefectCoop|60\%Coop}$, $a_{Defect|60\%Defect} = g_{Defect|60\%Defect}$, and $a_{Defect|60\%Coop} = g_{Defect|60\%Coop}$). Base Model 1 had six degrees of freedom and fit the data well, $G^2(6) = 4.43$, $p = .62$.

First, we tested whether old–new discrimination was affected by opponent type or group by setting all parameters to be equal that represent the probability of recognizing a face from the trust game as old (i.e., $D_{Defect|60\%Defect} = D_{Coop|60\%Defect} = [D_{Irrelevant|60\%Defect} = D_{New|60\%Defect}] = D_{Defect|60\%Coop} = D_{Coop|60\%Coop} = [D_{Irrelevant|60\%Coop} = D_{New|60\%Coop}]$). The additional restrictions were compatible with the data, $\Delta G^2(5) = 8.43$, $p = .13$, forcing us to conclude that old–new discrimination did not differ as a function of opponent type or group. This is consistent with the analysis of the old–new discrimination scores reported above.

Subsequently, we tested whether the probability of guessing that a face belonged to a defector differed between the two groups, by imposing, on Base Model 1, the restriction that $[a_{Defect|60\%Defect} = g_{Defect|60\%Defect}] = [a_{Defect|60\%Coop} = g_{Defect|60\%Coop}]$. The restriction was clearly incompatible with the data, $\Delta G^2(1) = 477.64$, $p < .001$

($[a_{Defect|60\%Defect} = g_{Defect|60\%Defect}] = .82$ [$CI$ = 0.79–0.85], and $[a_{Defect|60\%Coop} = g_{Defect|60\%Coop}] = .27$ [$CI$ = 0.24–0.30]). The values of the parameters correspond approximately to the actual relative proportion of defector versus cooperator faces in the 60% defectors group and the 60% cooperators group (75% and 25% respectively), which is consistent with probability matching (see Spaniol & Bayen, 2002). If anything, the guessing biases were even more extreme than the game-phase ratios of defectors to cooperators would have warranted. Thus, the analysis of the guessing biases show that the ratio of defectors and cooperators was represented, which is a necessary precondition for analyzing the effects of the ratio manipulation on memory.

Next, we tested whether source memory for defectors and cooperators differed between the 60% defectors group and the 60% cooperators group. The restrictions that source memory for defectors and cooperators did not differ between the groups, $d_{Defect|60\%Defect} = d_{Defect|60\%Coop}$, and $d_{Coop|60\%Defect} = d_{Coop|60\%Coop}$, were compatible with the data, $\Delta G^2(2) = 1.22$, $p = .54$, forcing us to conclude that there were no differences in source memory between the two groups. Source memory for the rare opponent types was not even descriptively better than source memory for the frequent opponent types (see Fig. 2). The restriction that the source memory parameters for irrelevant faces did not differ between groups, $d_{Irrelevant|60\%Defect} = d_{Irrelevant|60\%Coop}$, was also compatible with the data, $\Delta G^2(1) = 0.79$, $p = .37$, suggesting that source memory for irrelevant behavior did not differ between the two groups. The restrictions implying that source memory does not differ between groups were incorporated into Base Model 2, which also fit the data well, $G^2(9) = 6.60$, $p = .68$.

To test whether source memory differed as a function of opponent type, we imposed, on Base Model 2, the restriction that $d_{Defect} = d_{Coop}$, which was compatible with the data, $\Delta G^2(1) = 0.16$, $p = .69$, suggesting that source memory did not differ between faces of defectors and cooperators. The restrictions applied so far were incorporated into Base Model 3, which also fit the data well, $G^2(10) = 6.76$, $p = .75$. Next, we tested whether source memory for irrelevant faces was worse than source memory for faces of defectors and faces of cooperators, by imposing, on Base Model 3, the restriction that $[d_{Defect} = d_{Coop}] = d_{Irrelevant}$. The restriction was incompatible with the data, $\Delta G^2(1) = 47.53$, $p < .001$. We therefore conclude that source memory for irrelevant faces was worse than source memory for defectors and cooperators. This replicates the source memory results of Experiment 1.

The analysis based on the multinomial model of source memory helps understanding the interaction effects between opponent type and group on the CSIM and the likability ratings that seemed to be inconsistent at first glance. The CSIM is known to confound source memory and guessing. It is thus not surprising to see that the CSIM was heavily influenced by the tendency to guess that a face of which the source was not known belonged to the most frequent opponent type. Observable source classifications in the CSIM's numerator (see Section 2.2.4 of Experiment 1) were disproportionately large for faces belonging to the most frequent opponent type because *both* source memory *and* guessing biases necessarily increased the number of correct classification of these faces. This may also explain why Singer et al. (2004) found a higher number of correct source classifications for faces of cooperators.

Similarly, the test-phase likability ratings most plausibly reflect the influence of both source memory *and* guessing. The likability ratings for cooperator and defector faces are in the direction that one would expect assuming that participants had some memory for the behavior that was associated with the defector and the cooperator faces—defector faces were rated less likable than cooperator faces in both groups. Interestingly, the likability ratings of the irrelevant faces were very similar to the ratings of the faces that belong to the opponent type that was most frequently encountered in the trust game. In the 60% defectors group, irrelevant faces were rated as unlikable as defector faces. In the 60% cooperators group, irrelevant faces were rated as likable as cooperator faces. To understand this pattern of results, it is important to bring to mind that participants had virtually no source memory at all for the irrelevant faces and thus could only guess the category of the faces. As was to be expected, they guessed most frequently that the faces belonged to the opponent type that occurred most frequently in the experiment, as a consequence of which the likability ratings were the same for the irrelevant faces and the faces of the most frequent category. Thus, this finding could be explained by assuming that participants tried to give consistent evaluations and therefore gave lower likability ratings when they guessed that a face belonged to a defector and higher likability ratings when they guessed that a face belonged to a cooperator. However, a more interesting possibility is that the tendency towards guessing that a face belonged to a defector or to a cooperator was accompanied by genuine emotional reactions towards the faces. From a functional perspective, it would make sense that participants' emotional reactions towards faces of which the source is not known are more negative in an environment that consists largely of cheaters and more positive in an environment that consists largely of cooperators. It is plausible that these emotional evaluations go hand in hand with approach and avoidance tendencies and may also increase or decrease the motivation to engage in social exchange. Thus, it is possible that likability can be considered a regulatory variable for social interaction (cf. Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008), which is consulted when making the decision whether to engage in social exchange with a particular individual or not.

### 3.3. Discussion

In Experiment 2, base rates of defectors and cooperators were manipulated to see whether it was possible to replicate Barclay's finding that rarity determines memory for cheaters and cooperators in the context of a social-dilemma game. In contrast to Barclay's study, the present study included a neutral control condition, which allowed us to assess the relative importance of rarity versus social relevance in determining memory. The results of Experiment 2 show clearly that memory was primarily determined by social relevance. The finding of Experiment 1 that source memory for faces of defectors and cooperators is better than source memory for irrelevant faces was replicated. In contrast, the effect of opponent type on old–new face recognition that was present in Experiment 1 was not replicated in Experiment 2. This fits to previous results showing that source memory seems to be more sensitive to the social relevance of the stimulus material than old–new discrimination (Bell & Buchner, 2010, in press-a, in press-b; Buchner et al., 2009).

Contrary to the findings of Barclay (2008), old–new discrimination and source memory were not affected by the ratio manipulation although, descriptively, old–new discrimination was better for the opponent type that was rare and also socially relevant. Source memory was not even descriptively worse for the more frequent opponent type. Source memory for the irrelevant opponents was clearly worse than source memory for defectors and cooperators although the irrelevant opponents were underrepresented relative to the socially relevant opponents in both groups, which should have led to an increase, not a decrease of source memory for irrelevant faces if rarity were the primary determinant of source memory. Thus, the results clearly demonstrate that the primary determinant of source memory is social relevance, and not rarity. This is consistent with experiments using the *description paradigm* showing that source memory for faces of cheaters was unrelated to the degree to which the cheating behavior was exceptional (Buchner et al., 2009).

The results of Experiment 2 are inconsistent with those obtained by Barclay (2008). Note that the discrepancy between the studies cannot be ascribed to a lack of statistical power, because sample size, and statistical power, were much larger in the present study ($N = 113$) than in Barclay's study ($N = 40$ in the two groups with a majority of defectors or cooperators). It would be possible to explain the discrepancy between studies by assuming that the rarity effects observed by Barclay were the consequence of a strategic effort to encode and remember the cheaters and cooperators that was enforced by the instructions used in his study. However, it is also possible to attribute this discrepancy to other methodological differences between studies. In the present Experiment 2, the rare opponent type comprised 20% of the faces, just as in Barclay's study, but we also included irrelevant control faces (20% of the faces). Assuming that participants ignored the irrelevant faces, the ratio of defectors to cooperators was slightly more extreme in Barclay's study (80% vs. 20% of the faces) than in the present Experiment 2 (75% vs. 25% of the exchange-relevant faces). It is also possible that the (small) influence of rarity may have been overridden by the large social-relevance effect observed in Experiment 2. To see whether this may explain why Barclay found a modulation of memory by rarity, we omitted the faces associated with irrelevant information in Experiment 3.

## 4. Experiment 3

### 4.1. Method

#### 4.1.1. Participants

Participants were 81 persons (62 women). Their age ranged from 18 to 48 years ($M = 23$). The participants were randomly assigned to two groups (see below).

#### 4.1.2. Materials and procedure

Materials and procedure were identical to Experiment 2 with the following exceptions. In the trust game, 60 randomly selected pictures were used. In the 80% defectors group, 48 opponents (80%) defected, and 12 opponents (20%) cooperated. In the 80% cooperators group, 48 opponents (80%) cooperated, and 12 opponents (20%) defected. The trust game was split up in four consecutive blocks of 15 trials that consisted of 12 trials of the frequent condition, and 3 trials of the rare condition. The order of trials within a block was randomly determined.

In the test phase, 48 facial photographs were presented. Twenty-four of these photographs were new, and 24 were old. Of these, 12 had been associated with defectors, and 12 had been associated with cooperators in the trust game.

#### 4.1.3. Design

The between-subjects independent variable was the proportion of defectors and cooperators in the trust game (*80% defectors*, *80% cooperators*) and the within-subject variable was opponent type (*defector*, *cooperator*). The dependent variables were test-phase likability ratings, old–new discrimination, and source judgments given an "old" judgment. Given a sample size of $N = 81$, and $\alpha = .05$, effects of size $w = 0.06$ could be detected for the interaction between group and opponent type with a probability of $1 - \beta = .95$.

### 4.2. Results

#### 4.2.1. Trust game investments

A $2 \times 2$ MANOVA with group and opponent type as independent variables showed that participants in the 80% defectors group invested less (23 cents) than participants in the 80% cooperators group (31 cents), $F(1, 79) = 17.54$, $p < .001$, $\eta^2 = .18$. As was to be expected, there was no significant main effect of opponent type and no interaction between the two variables. Investments in the 80% defectors group were significantly lower than investments in Experiment 1, $t(127) = 3.99$, $p < .001$. Investments in the 80% cooperators group were only descriptively enhanced in comparison to Experiment 1, $t(128) = 0.90$, $p = .37$. This is consistent with the findings of Experiments 1 and 2 in which participants also pursued a cautious investment strategy. Nevertheless, the significant effect of group on investments showed that the proportion of defectors versus cooperators affected the willingness to engage in social exchange.

#### 4.2.2. Old–new discrimination

Old–new discrimination (Table 1) did not differ as a function of group, $F(1, 79) = 0.08$, $p = .78$, $\eta^2 < .01$, or opponent type, $F(1, 79) = 2.07$, $p = .16$, $\eta^2 = .03$. However, there was a significant interaction between the two variables, $F(1, 79) = 4.44$, $p = .04$, $\eta^2 = .05$. Cooperator faces were better recognized than defector faces in the 80% defectors group, but defector faces were better recognized than cooperator faces in the 80% cooperators group.

#### 4.2.3. Test-phase likability ratings

Test-phase likability ratings (Table 1) did not differ between groups, $F(1, 79) = 1.04$, $p = .31$, $\eta^2 = .01$. There was a significant main effect of opponent type, $F(1, 79) = 31.62$, $p < .001$, $\eta^2 = .29$. Defector faces were less likable than cooperator faces. There was no significant interaction between the two variables, $F(1, 79) = 0.62$, $p = .44$, $\eta^2 < .01$.

*4.2.4. Source memory*

The analysis of the CSIM (Table 1) revealed a significant interaction between opponent type and group, $F(1, 79) = 116.54$, $p < .001$, $\eta^2 = .60$, indicating more accurate conditionalized source classifications for the more frequent opponent type. The source memory data were also analyzed using the multinomial model proposed by Bayen et al. (1996; see Fig. 4). The model is identical to the source memory model used in Experiments 1 and 2 with the only exception that there are only three types of items (defectors, cooperators, and new faces), and three types of responses ("defector", "cooperator", "new"). Given that there was no irrelevant condition in the present experiment, we assumed that the probability of recognizing a new face as "new" would be equal to the mean probability of recognizing an old face as "old" ($D_{New} = [D_{Defect} + D_{Coop}]/2$), and we further assumed that the probability of guessing that a recognized face belongs to a defector would be equal to the probability of guessing that a non-recognized face belonged to a defector ($a_{Defect} = g_{Defect}$) in both groups in order to obtain an identifiable model. Based on the results of the old–new discrimination reported above showing no significant effect of opponent type and no significant effect of group, but a significant interaction between opponent type and group, we decided to start our analysis by setting the parameters that represent the probability of recognizing a face of the frequent opponent type to be equal across groups (i.e., $D_{Defect|80\%Defect} = D_{Coop|80\%Coop}$) and by setting the parameters that represent the probability of recognizing a face of the rare opponent type to be equal across groups (i.e., $D_{Coop|80\%Defect} = D_{Defect|80\%Coop}$). The resulting base model (henceforth Base Model 1) fit the data well, $G^2(2) = 2.06$, $p = .36$. In a next step, we tested whether the old–new discrimination parameters differed between opponent types of the frequent category and opponent types of the rare category by imposing, on Base Model 1, the restriction that $[D_{Defect|80\%Defect} = D_{Coop|80\%Coop}] = [D_{Coop|80\%Defect} = D_{Defect|80\%Coop}]$. The restriction was incompatible with the data, $\Delta G^2(1) = 4.19$, $p < .05$. Old–new discrimination for faces of opponents of the frequent category was worse than old–new discrimination for faces of opponents of the rare category ($[D_{Defect|80\%Defect} = D_{Coop|80\%Coop}] = .59$ [$CI = 0.55–0.63$]; $[D_{Coop|80\%Defect} = D_{Defect|80\%Coop}] = .66$ [$CI = 0.62–0.69$]). This mirrors the results of the old–new discrimination reported above.

The interesting question was whether we would find the same pattern of results for the source memory parameters. We started by imposing, on Base Model 1, the restrictions that source memory for opponents of the frequent category would be equal across groups (i.e., $d_{Defect|80\%Defect} = d_{Coop|80\%Coop}$) and that source memory for opponents of the rare category would be equal across groups (i.e., $d_{Coop|80\%Defect} = d_{Defect|80\%Coop}$). The restrictions were compatible with the data, $\Delta G^2(2) = 1.51$, $p = .47$, suggesting that source memory for opponents of the frequent category did not differ between groups, and that the same was true for source memory for opponents of the rare category (i.e., $d_{Defect|80\%Defect} = d_{Coop|80\%Coop}$, and $d_{Coop|80\%Defect} = d_{Defect|80\%Coop}$, respectively). The restrictions applied so far were incorporated into Base Model 2, which, as expected, also fit the data well, $G^2(4) = 3.57$, $p = .46$. In a next step, we tested whether source memory for opponents of the frequent category would be worse than source memory for opponents of the rare category. Descriptively, this seemed to be the case (see Fig. 5). The null hypothesis of

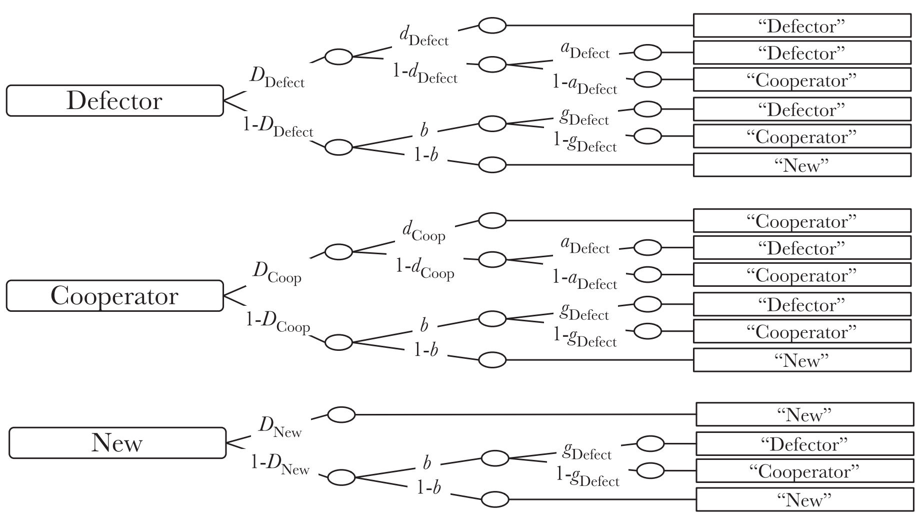


**Fig. 4.** Bayen et al.'s (1996) source memory model as adapted for Experiment 3. Rounded rectangles on the left side represent the types of faces presented (*defectors*, *cooperators*, *new faces*). Letters along the links represent the probabilities with which certain cognitive states occur ($D$: probability of correctly identifying a face as old or new; $d$: source memory in the sense of remembering the context of encountering a face that was detected as old; $b$: probability of guessing that a non-recognized face is old; $a$: probability of guessing that a recognized face was encountered in a particular context; $g$: probability of guessing that a non-recognized face that was guessed to be old was encountered in a particular context). Rectangles on the right side represent the categories of participant's judgments.

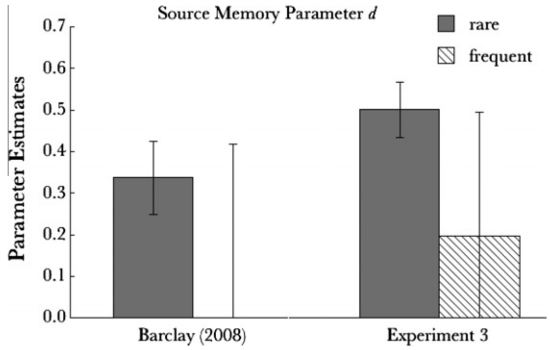


**Fig. 5.** Parameter estimates for the source memory parameters for faces associated with rare opponent strategies [$d_{Defect|80\%Coop} = d_{Coop|80\%Defect}$] and for faces associated with frequent opponent strategies [$d_{Defect|80\%Defect} = d_{Coop|80\%Coop}$] in Experiment 3. The parameters represent conditional probabilities of correct source identifications given correct old–new discrimination. Error bars represent the .95 confidence intervals.

no such difference is tested by imposing, on Base Model 2, the restriction that source memory for opponents of the frequent type is equal to source memory for opponents of the rare type, [$d_{Defect|80\%Defect} = d_{Coop|80\%Coop}$] = [$d_{Coop|80\%Defect} = d_{Defect|80\%Coop}$]. The restriction was incompatible with the data, $\Delta G^2(1) = 4.12$, $p < .05$, forcing us to conclude that source memory for the frequent opponent types (i.e., $d_{Defect|80\%Defect}$ and $d_{Coop|80\%Coop}$ combined) was worse than source memory for the rare opponent types (i.e., $d_{Coop|80\%Defect}$ and $d_{Defect|80\%Coop}$ combined).

The probability of guessing that a face (of which the source was not known) belonged to a defector was again clearly different between the two groups, $\Delta G^2(1) = 125.18$, $p < .001$ ([$a_{Defect|80\%Defect} = g_{Defect|80\%Defect}$] = .88 [$CI$ = 0.82–0.93]; and [$a_{Defect|80\%Coop} = g_{Defect|80\%Coop}$] = .27 [$CI$ = 0.20–0.33]). This shows that the ratio of defectors to cooperators was cognitively represented.

#### *4.2.5. Reanalysis of the source monitoring results of Barclay (2008)*

So far, the results of the present Experiment 3 seem generally consistent with the results of Barclay (2008). Given that Barclay used a different approach to analyzing the source monitoring data, it is interesting to see whether a reanalysis of his data using the source memory model depicted in Fig. 4 would yield the same conclusions. The procedure in that experiment differed somewhat from the one used here. For instance, Barclay asked his participants to classify the faces into "defectors" and "cooperators" even when they had previously classified the face as being "new". Furthermore, there was a third group in which participants encountered 50% defector faces and 50% cooperator faces. Thus, to compare the results of the present Experiment 3 and Barclay's results directly, we have to ignore certain aspects of his data (i.e., the source classifications for faces previously classified as "new", and the results in the 50% defectors group). Base Model 1 fit the remaining data well, $G^2(2) = 3.72$, $p = .15$. The restriction that old–new discrimination for faces associated with the frequent opponent type is equal to old–new discrimination for faces associated with the rare opponent type was incompatible with the data, $\Delta G^2(1) = 11.24$, $p < .001$. This mirrors the old–new discrimination results reported by Barclay. Setting the source memory parameters for opponents of the frequent type and for opponents of the rare type to be equal across groups (i.e., $d_{Defect|80\%Defect} = d_{Coop|80\%Coop}$, and $d_{Coop|80\%Defect} = d_{Defect|80\%Coop}$) is compatible with the data, $\Delta G^2(2) = 0.40$, $p = .82$. Accordingly, Base Model 2 fit the data very well, $G^2(4) = 4.12$, $p = .39$. Setting the source memory parameters for opponents of the frequent type (i.e., $d_{Defect|80\%Defect}$ and $d_{Coop|80\%Coop}$ combined) and for opponents of the rare type (i.e., $d_{Coop|80\%Defect}$ and $d_{Defect|80\%Coop}$ combined) to be equal is incompatible with the data, $\Delta G^2(1) = 6.25$, $p < .01$ (see Fig. 5). In sum, our reanalysis confirms Barclay's conclusions.

### *4.3. Discussion*

The results of the source memory test confirmed our suspicion that rarity affects memory when there are only cooperators and defectors, and when the defectors-to-cooperators ratio is extreme. Both old–new discrimination and source memory were better for the rare opponent type than for the frequent opponent type. The results replicate the findings of Barclay (2008). Together with the results of Experiment 2, the present results provide empirical support for Barclay's preferred interpretation of his results that memory is determined by a mechanism that focuses on socially relevant behavior (defection, cooperation) in exchange situations and that flexibly adapts to the relative frequency of the behavioral strategies. Given the results of Experiment 2, we add to this interpretation that people seem to have a disposition to attend to both cooperators and defectors about equally unless the differences in base rates are really extreme. Nevertheless, memorizing the rare category may be associated with a high fitness value because one would only have to change one's dominant behavior strategy (e.g., to engage in social exchange in an environment that is dominated by cooperators) for the rare opponent type (e.g., for defectors in an environment that is dominated by cooperators; see also Aktipis, 2006; Barclay, 2008). Consistent with the assumption that guessing biases can be very useful in real life (e.g., Bayen, Nakamura, Dupuis, & Yang, 2000; Spaniol & Bayen, 2002), the present results show that a response strategy such as "if in doubt, guess that most faces belong to the frequent opponent type" may be very efficient when the ratio of frequent to rare opponents is extreme. This is evident from the fact that the CSIM was *most accurate* for the frequent opponent type *even though* source memory was extremely low. Thus, it may simply be uneconomic to store source information for the frequent opponent type for each face separately. A bias towards remembering the exceptions from this rule may complement this efficient response strategy and may further increase source classification performance.

## 5. General discussion

The present results can be summarized as follows. (1) In contrast to previous studies examining source memory for cheaters in the *description paradigm* (Bell & Buchner, in press-a; Buchner et al., 2009), defectors and cooperators were equally well remembered in the present *involvement paradigm*. (2) Old–new discrimination and source memory were better for defector and cooperator faces than for irrelevant control faces. (3) Frequency of opponent type modulated old–new discrimination and source memory to some degree, but only if (a) defectors and cooperators were contrasted directly and (b) the ratio of defectors to cooperators was rather extreme.

The most important finding of the present series of experiments is that source memory for faces of both defectors and cooperators is enhanced. This supports the hypothesis derived from theories of reciprocal altruism that information that is relevant for social exchange should be especially well remembered (Cosmides & Tooby, 1992). According to models of direct reciprocity (Axelrod & Hamilton, 1981; Trivers, 1971), cooperation is not evolutionary stable unless it is made contingent on the cooperation of others. Therefore, these models lead to the prediction that people should remember both defectors and cooperators. Source memory for an opponent's behavior may be adaptive because it allows one to alter one's behavior according to the opponent's behavior in previous encounters.

It seems plausible that the relative importance of cheating versus cooperation depends on the circumstances. There may be situations in which it is more beneficial to focus on cheaters and situations in which it is more beneficial to focus on cooperators. For instance, when cooperation is the rule, it may be beneficial to remember the cheaters and to engage in social exchange with everyone else. When cheating is common, it may be a more successful strategy to remember the cooperators and to treat everyone else with suspicion (Aktipis, 2006; Barclay, 2008). It is highly plausible that the benefit associated with mechanisms that focus on all socially relevant information available in the environment and that flexibly adjust to the relevance of this information in a specific situation (e.g., the social environment or the internal emotional state) would be much higher than that of a mechanism that is restricted to the processing of only a small part of the available information (i.e., cheating).

The present results also show that people seem to focus on both cooperation and defection unless the ratio of cooperators to defectors is extreme. The modulation of memory by rarity was small and unreliable in comparison to the large social-relevance effect. This finding is important because it refutes a domain-general rarity explanation that could have been applied to explain the results of previous studies examining source memory for cheaters. Barclay (2008) did not test whether memory for exchange-related information was enhanced in comparison to other information. Given that there is a large body of evidence showing that rare or unusual events are better attended (e.g., Parmentier et al., 2008; Schröger & Wolff, 1998) and remembered (Hunt & Worthen, 2006; Schmidt, 1991), a simple and plausible conclusion from his results is that memory for cooperators and defectors was determined by a domain-general rarity detection system, or a "general tendency to focus on rarity and remember exceptions and atypical stimuli" (Barclay, 2008, p. 819).

However, the results of Experiment 2 are clearly inconsistent with the assumption that memory is primarily determined by rarity. In this experiment, social relevance had a considerable effect on source memory, whereas rarity had no effect on memory at all. There was hardly any source memory for irrelevant control faces even though these faces were rare. This shows that source memory is primarily determined by social relevance, not rarity. Rarity seems to modulate memory only when base rate differences are extreme and when participants are not distracted by irrelevant information. Together, these findings suggest that the effect of rarity on memory for cooperators and defectors should not be overstated. This conclusion is consistent with results from Buchner et al. (2009), who found that the source memory advantage for faces that were associated with descriptions of cheating was not affected by the exceptionality of the behavior descriptions.

An interesting question is why the selective source memory advantage for cheaters observed previously under a variety of conditions (Bell & Buchner, in press-a, in press-b; Buchner et al., 2009) was replaced by an advantage of both defectors and cooperators over irrelevant persons in the present series of experiments. Most plausibly the discrepancy is related to the fact that participants were directly involved into social exchange with the characters in the present experiments (i.e., suffered negative consequences from interactions with cheaters and benefitted from interactions with cooperators), whereas they only read about interactions involving third parties in those previous experiments. It has been shown that cheater detection in the Wason selection task is sensitive to perspective effects (Gigerenzer & Hug, 1992) in that cheater detection is more pronounced if a participant is cued into the role of a potential victim of cheating than when he or she is cued into the role of the cheater. The present findings suggest that memory for cheaters may also be sensitive to perspective effects in that the relative significance of cheating and cooperation may change depending on whether one is cued into the role of a social exchange partner or into the role of a witness of social exchange. In the *description paradigm*, there is no need to memorize whom to reciprocate which is why we can ignore non-cheaters, but we should still attend to norm violations of others. This interpretation is consistent with empirical studies and theorizing about social norms which focus on third party condemnation (DeScioli & Kurzban, 2009), and third-party punishment (Fehr & Fischbacher, 2004a, 2004b) of failures to reciprocate. The results in the *involvement paradigm* may differ from those in the *description paradigm* for at least two reasons. First, the *involvement paradigm* may strongly prime the possibility of future encounters with the opponents, raising memory performance for cooperators with whom we want to maintain the already established cooperative relationship. Furthermore, we may have to remember people from whom we received a benefit because we have the obligation to reciprocate this favor in future interactions. Speculatively, experiencing cooperation directly may also lead to more intense positive emotions (relief, gratitude, joy) than merely reading about cooperative encounters between third parties, which might result in better memory for cooperative encounters in the *involvement* paradigm relative to the *description* paradigm.

In summary, the present results show that old–new discrimination and source memory are better for faces of defectors and cooperators than for irrelevant faces. Experiment 3 replicates the finding of Barclay (2008) that old–new discrimination and source memory for faces of defectors and cooperators is modulated by the frequency of the behavior. A mechanism that rigidly focuses on cheating while ignoring other relevant information may indeed be inferior to a mechanism that focuses on socially relevant behavior and is capable to flexibly adapting to the frequency of behavior in the population.

## Acknowledgment

We thank G. Schröer, E. Winnebeck, and A. Könnecke for their help during data acquisition. We also thank Leda Cosmides and three anonymous reviewers for their helpful comments on a previous version of the manuscript. Finally, thanks are due to Pat Barclay for giving us the opportunity to reanalyze his data. The research reported in this article was supported by a grant from the Deutsche Forschungsgemeinschaft.

## References

Aktipis, C. A. (2006). Recognition memory and the evolution of cooperation: How simple strategies succeed in an agent-based world. *Adaptive Behavior, 14*, 239–247.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*, 1390–1396.

Barclay, P. (2008). Enhanced recognition of defectors depends on their rarity. *Cognition, 107*, 817–828.

Barclay, P., & Lalumière, M. L. (2006). Do people differentially remember cheaters? *Human Nature, 17*, 98–113.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97*, 548–564.

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 197–215.

Bayen, U. J., Nakamura, G. V., Dupuis, S. E., & Yang, C. L. (2000). The use of schematic knowledge about sources in source monitoring. *Memory & Cognition, 28*, 480–500.

Bell, R., & Buchner, A. (in press-a). Justice sensitivity and source memory for cheaters. *Journal of Research in Personality.*

Bell, R., & Buchner, A. (in press-b). Source memory for faces is determined by their emotional evaluation. *Emotion.*

Bell, R., & Buchner, A. (2010). Valence modulates source memory for faces. *Memory & Cognition, 38*, 29–41.

Bröder, A., & Meiser, T. (2007). Measuring source memory. *Zeitschrift fur Psychologie/Journal of Psychology, 215*, 52–60.

Buchner, A., Bell, R., Mehl, B., & Musch, J. (2009). No enhanced recognition memory, but better source memory for faces of cheaters. *Evolution and Human Behavior, 30*, 212–224.

Chiappe, D., Brown, A., Dow, B., Koontz, J., Rodriguez, M., & McCulloch, K. (2004). Cheaters are looked at longer and remembered better than cooperators in social exchange situations. *Evolutionary Psychology, 2*, 108–120.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187–276.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York, NY, US: Oxford University Press.

Cosmides, L., & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 584–627). Hoboken, NJ, US: John Wiley & Sons Inc.

D'Argembeau, A., & Van der Linden, M. (2004). Influence of affective meaning on memory for contextual information. *Emotion, 4*, 173–188.

DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition, 112*, 281–299.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*, 187–190.

Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior, 25*, 63–87.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43*, 127–171.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*, 546–567.

Hammerl, M. (2000). *Der Operationalisierungseffekt. Über den Einfluss der Forschungsstrategie auf die Befunde der experimentellen Psychologie. [The operationalization effect. How the research strategy determines findings in experimental psychology.]*. Lengerich: Pabst Science Publishers.

Hunt, R. R., & Worthen, J. B. (2006). *Distinctiveness and memory*. New York: Oxford University Press.

Klein, S. B., Cosmides, L., Tooby, J., & Chance, S. (2002). Decisions and the evolution of memory: Multiple systems, multiple functions. *Psychological Review, 109*, 306–329.

Mealey, L., Daood, C., & Krage, M. (1996). Enhanced memory for faces of cheaters. *Ethology and Sociobiology, 17*, 119–128.

Mehl, B., & Buchner, A. (2008). No enhanced memory for faces of cheaters. *Evolution and Human Behavior, 29*, 35–41.

Moshagen, M. (2010). MultiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*, 42–54.

Nairne, J. S. (2005). The Functionalist Agenda in Memory Research. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 115–126). Washington, DC, US: American Psychological Association.

Nairne, J. S., & Pandeirada, J. N. S. (2008). Adaptive memory: Remembering with a stone-age brain. *Current Directions in Psychological Science, 17*, 239–243.

Oda, R. (1997). Biased face recognition in the prisoner's dilemma game. *Evolution and Human Behavior, 18*, 309–315.

O'Gorman, R., Wilson, D. S., & Miller, R. R. (2008). An evolved cognitive bias for social norms. *Evolution and Human Behavior, 29*, 71–78.

Parmentier, F. B. R., Elford, G., Escera, C., Andres, P., & San Miguel, I. (2008). The cognitive locus of distraction by acoustic novelty in the cross-modal oddball task. *Cognition, 106*, 408–432.

Schmidt, S. R. (1991). Can we have a distinctive theory of memory? *Memory & cognition, 19*, 523–542.

Schröger, E., & Wolff, C. (1998). Behavioral and electrophysiological effects of task-irrelevant sound change: A new distraction paradigm. *Cognitive Brain Research, 7*, 71–87.

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

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34–50.

Spaniol, J., & Bayen, U. J. (2002). When is schematic knowledge used in source monitoring? *Journal of Experimental Psychology: Learning, Memory and Cognition, 28*, 631–651.

Tooby, J., Cosmides, L., Sell, A., Lieberman, D., & Sznycer, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. In A. J. Elliot (Ed.), *Handbook of approach and avoidance motivation* (pp. 251–271). New York, NY: Psychology Press.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*, 35–57.