The effect of social categorization on trust decisions in a trust game paradigm

Elena Cañadas [1]*, Rosa Rodríguez-Bailón [2] and Juan Lupiáñez [3]

[1] Department of Organizational Behavior, University of Lausanne, Lausanne, Switzerland, [2] Department of Social Psychology, Mind, Brain and Behavior Research Center, University of Granada, Granada, Spain, [3] Department of Experimental Psychology, Mind, Brain and Behavior Research Center, University of Granada, Granada, Spain

This study investigates whether participants use categorical or individual knowledge about others in order to make cooperative decisions in an adaptation of the trust game paradigm. Concretely, participants had to choose whether to cooperate or not with black and white unknown partners as a function of expected partners' reciprocity rates. Reciprocity rates were manipulated by associating three out of four members of an ethnic group (blacks or whites consistent members) with high (or low) reciprocity rates, while the remaining member of the ethnic group is associated with the reciprocity of the other ethnic group (inconsistent member). Results show opposite performance's patterns for white and black partners. Participants seemed to categorize white partners, by making the same cooperation decision with all the partners, that is, they cooperated equally with consistent and inconsistent white partners. However, this effect was not found for black partners, suggesting a tendency to individuate them. Results are discussed in light of the implications of these categorization-individuation processes for intergroup relations and cooperative economic behavior.

Keywords: trust game, categorization, individuation, ingroup–outgroup perception, cooperation

INTRODUCTION

Every day we come across many people. The amount of information that we can extract from these encounters can be so demanding that it needs to be organized in order to be used for making efficient decisions and plan our subsequent actions toward those people. Such organization of the information provides us with general knowledge about the individuals we intend to interact with. At the same time, this knowledge helps us guiding our interactions even with strangers.

When perceiving a person for the first time we may categorize him/her according to the salient features of his/her face such as sex, age, or ethnicity (Devine, 1989; Fiske and Neuberg, 1990). Research shows that people use stereotypes to attribute characteristics to others and consequently the impressions we form about them can be biased by those stereotypes. Interestingly, this process can take place outside the individual's awareness (Bargh and Williams, 2006; Cunningham and Zelazo, 2007).

Social perception may involve a decision-making process where social agents decide whom to interact with and how. Perceivers try to predict the course of the interaction and whether the goals of the interaction will be achieved or not. In social contexts, this decision-making process is influenced by certain salient features of the people we interact with, such as facial expressions (e.g., Scharlemann et al., 2001; Ruz and Tudela, 2011), physical attractiveness (e.g., Solnick and Schweitzer, 1999; Solnick, 2001), or ethnicity (e.g., Sommers, 2006, 2007), which may influence our beliefs and

expectations about those with whom we have to interact (e.g., Ruz et al., 2011; Gaertig et al., 2012), especially when we know nothing about them. One of the most crucial features when interacting with others concerns the level of trust deployed in these relationships. Trust is essential for a secure and healthy social life (Dunning et al., 2014), being considered as a core social motive (Fiske, 2003).

Although essential to social life, trust is conceived as irrational by philosophers (e.g., Hobbes, 1997; Machiavelli, 1515/2003) or neoclassical economists (e.g., Berg et al., 1995; Bolle, 1998). Despite that, empirical evidence has shown that people trust strangers and reward that trust (for reviews, see Johnson and Mislin, 2011; Balliet and Van Lange, 2013).

Outside the lab, trust is present in interpersonal situations (trusting a confidant), in economic markets (trusting a financial advisor), or even in political elections (trusting a government). Knowing whom to trust is crucial for preventing being deceived by others, being taken advantage of, or avoiding financial losses, and many other undesirable outcomes. When we have previous experience with our partner, the object of our trust, we can predict at different levels of certainty whether we can trust him/her, and in fact minimal interactions can already influence trustworthiness judgments (Frank et al., 1993). However, when we lack this previous experience with somebody it is difficult (although not impossible) to predict his/her behavior and consequently trust or not him/her[1]. However, trust at zero acquaintance has to be influenced by factors different from the experience with the trustee (Dunning et al., 2014).

Some research has focused on some of these factors that may influence participants trusting behavior. Among these, there is an emerging literature pointing to the role of shared group membership in the promotion of trust (Platow et al., 2012). Much of this work follow the theoretical claim by Brewer (1981, p. 356) that group membership "serves as a rule for defining the boundaries of low-risk interpersonal trust that bypasses the need for personal knowledge and the costs of negotiating reciprocity." This proposition has been supported by the results of some studies showing that participants trust others as a function of their group membership (e.g., Tanis and Postmes, 2005), although it is not supported by results from other studies, as we will describe later (Tortosa et al., 2013).

As in many other impression formation processes, when deciding to trust unfamiliar others, we can either: (a) categorize them and interact with them according to the inferences that can be extracted from what we know (or have learnt) about their category (i.e., their inferred group membership), or (b) to individuate them and try to predict their behavior in order to know how to interact with them based solely on what we specifically learnt about them. Literature has repeatedly shown that categorization seem to be the default process in particular for social stimuli (Brewer, 1988; Fiske and Neuberg, 1990; Kawakami et al., 1998; Cuddy et al., 2004; Nelson, 2005).

In the present study we were interested in evaluating whether people infer information (i.e., reciprocation rate) about others

[1]Trusting behavior is the willingness of an individual to expose themselves to the actions of others while trustworthy behavior is defined to be rewarding trust through reciprocation.

based primarily on their category membership (i.e., ethnicity) or on individuation perception, and consequentially their decision to trust them (e.g., sharing money) depends on these reciprocation inferences. In order to do so, we adapted a procedure developed to investigate the use of social categories for the allocation of attentional control (Canadas et al., 2013), to investigate the categorization-individuation processes underlying the cooperation dynamics in a trust game context.

Canadas et al. (2013) procedure presented photographs of men and women as the context in which congruent or incongruent stimuli appeared for participants to solve a flanker task. Three faces in each social category (i.e., consistent faces) were associated either with a high proportion of congruent trials (75% congruent–25% incongruent, i.e., a low conflict context) or a low proportion of congruent trials (25% congruent–75% incongruent, i.e., a high conflict context). Whereas a forth face in each group (i.e., inconsistent face) was associated to the proportion congruent of the other group. The extent to which inconsistent faces produced the same pattern of results as consistent ones, in spite of being associated to the opposite proportion of congruency as the social category they belonged to, was taken as an index of social categorization. A categorization pattern was observed in fact in the first study, thus supporting the abovementioned idea that categorization processes seem to be the default for social stimuli.

In a second study, we manipulated the instructions given to participants, either to individuate (i.e., pay attention to the individual characteristics of the faces) or categorize (i.e., pay attention to the category-based features of the faces). The previous pattern of results was replicated in the categorization instructions group. However, the pattern of results in the individuation instructions group showed that different effects were observed for consistent and inconsistent faces, thus reflecting the individual association rather than the group associations between faces and the proportion of congruency. This pattern was taken as evidence for individuation even in social contexts when participants are motivated to do so.

In conclusion, although categorization seems to play a dominant role in person perception processing, a wide range of variables has been shown to function as modulators of categorical thinking activation, including instructions, motivation, goals, and strategies (e.g., Lepore and Brown, 1997; Castelli et al., 2004; Macrae and Cloutier, 2009); the paper by Canadas et al. (2013) showed to be a suitable method to investigate these processes. In the current study we aimed at extending this procedure to investigate individuation-categorization processes in a more direct and clear social behavior, the decisions about to trust or not a partner in an economic game.

Wilson and Kayatani (1968) examined the effect of a partner's ethnicity on cooperation behavior in a prisoner's dilemma game. They found that participants were far more cooperative with ingroup partners than with outgroup partners. Also, Chen et al. (2010) performed several experiments using the prisoner's dilemma task in which participants identify themselves as members of the same university of the same ethnicity, and showed the same pattern of results as Wilson and Kayatani (1968), that is, the sense of belonging to the same group played an important role

in the participants' cooperation ratings. However, Tortosa et al. (2013) did not find evidence for this bias against the outgroup with the well-known Trust Game paradigm developed by Berg et al. (1995), originally constructed by Camerer and Weigelt (1988). In a first experiment Tortosa et al. (2013) observed no effect of ethnicity on the cooperation rate, whereas in the second experiment they observed in fact a smaller but reliable tendency toward a larger cooperation rate for the outgroup partners (64.4%) compared to the ingroup ones (57.5%).

In the current research we used a modified version of this trust game procedure and incorporated the Cañadas et al.'s (2013) manipulation to investigate categorization and individuation processes. We evaluated whether participants prefer to trust (cooperate/share money with) ingroup members (i.e., white partners), compared to outgroup members (black partners). The task typically involves two players, a trustor and a trustee. The trustor (participant) is endowed with a sum of money and has to decide whether or not to share it with her/his game partner. If s/he decides to keep the money for her/himself, the trustee gets nothing. If s/he decides to share the money, the trustee receives the initial endowment multiplied by an amount (determined by the experimenter). If the trustee then reciprocates, the sum is divided between the two players; otherwise the trustor obtains nothing. In this game, the typical decision of the trustor is hazardous because the trustee's reciprocation is not enforced by the rules. Still, substantial amounts of trust are observed across studies (Berg et al., 1995). These effects are attributed more to "social preferences" such as fairness, altruism, and reciprocation (see, for example; Fehr and Schmidt, 1999; Charness and Rabin, 2002; Fehr and Fischbacher, 2002, 2003; Fehr and Camerer, 2007) than to self-interest rational choices.

Importantly in our adaptation of the procedure, each participant was presented with two categories of faces (i.e., blacks and whites) of supposed partners randomly assigned to a high (75%) or low (25%) proportion of reciprocation in a within subject design. Also, as in Cañadas et al. (2013) we manipulated that one individual in each group (inconsistent member) is associated with the proportion of reciprocation of the other group. This will allow us to examine different effects of impression formation (participant's cooperation bias). Another advantage of using a within subject design is that it allows us to explore the learning processes underlying participants' strategy to adapt their sharing behavior with high reciprocation vs. low reciprocation partners.

A second and more important aim of our study is to evaluate whether participants individuate, or rather categorize, that is, the extent to which participant behave in the same way with all category members, irrespectively of whether they show a consistent or inconsistent cooperation rate with rest of the category members. In case of categorization, the same decision (e.g., to cooperate with the members of one ethnic group) will be displayed also with inconsistent members of the group (that is, even with those whose reciprocation rates are opposite to the one of their own ethnic-category, and equal to the other ethnic-category). On the contrary, participants will individuate to the extent that their decisions are taken accordingly to the reciprocation rate associated to each individual

face rather than to the ethnic-category. Therefore, in case of individuation, inconsistent individuals will be show different cooperation patterns than consistent faces.

We expected that along the block of trials, participants would use the facial features to categorize individuals according to the more salient features of their faces (i.e., ethnic features) and therefore decide to cooperate or not with them depending on the likelihood of their group to reciprocate. Thus, participants would share in greater extent with the individuals of the group more likely to reciprocate. However, we expected this to happen mainly for consistent individuals.

A different prediction was made for inconsistent individuals. On the one hand, and according to our previous research by Cañadas et al. (2013), inconsistent individuals might be also categorized. However, given the nature of the task (Trust Game round) when participants interact several times with the same partner, previous interaction with the same person influences the participant's decision to cooperate (King-Casas et al., 2005). Foregoing research has demonstrated that people attempting to maximize their benefits should learn from the feedback displayed after the interactions with the environment (reinforcement learning—Sutton and Barto, 1998) and consequently in our study individuation is a more efficient strategy. This strategy then should be learned quickly after the feedback of each interaction (Axelrod and Hamilton, 1981).

Taken all together, both the individualistic nature of this task and the explicit consequences of each decision (participants were informed about whether the partner reciprocated or not in each trial), we expected that participant would pay attention to each individual and therefore would individuate inconsistent partners, updating first impressions based on previous interactions (Chang et al., 2010; Campellone and Kring, 2013). This individuation pattern (i.e., a correlation between the participants cooperation rate and the individual reciprocation rate, nor the group reciprocation rate) was expected nevertheless mainly for the ingroup individuals (i.e., white partners), thus supporting previous knowledge on outgroup vs. ingroup social categorization (Judd and Park, 1988; Linville et al., 1989; Levin, 1996, 2000; MacLin and Malpass, 2001).
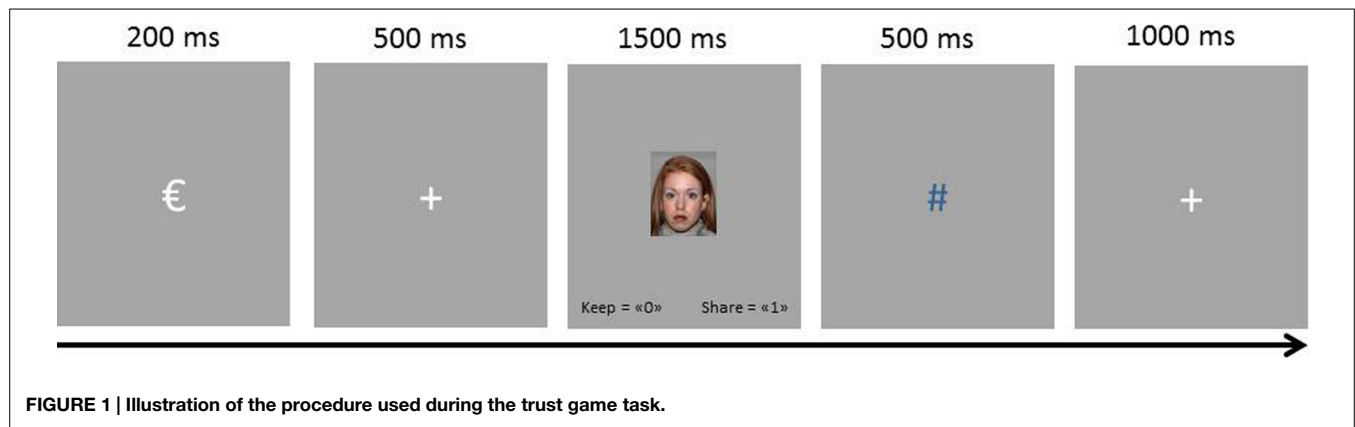
MATERIALS AND METHODS

Participants
Twenty-six undergraduate white students from the University of Granada (one man, mean age 20.15 years, SD = 1.93) participated in exchange for course credits. The study was conformed to the relevant regulatory standards approved by the local ethics committee of the University of Granada in the Department of Experimental Psychology. Participants signed consent forms and received 1% of the final payoffs (maximum 10 euros).

Stimuli and Procedure
At the beginning of the session, participants were instructed that the experiment explored the cooperation patterns that emerge between people during the so-called *trust game*. During the task participants played the role of "*trustors*." They received 1 euro and

FIGURE 1 | Illustration of the procedure used during the trust game task.

had to decide whether to keep or share it with an allegedly partner (i.e., the "*trustee*"), an unknown person for the participant from whom a picture is shown. Each trail starts with the Euro's symbol (€)—indicating that he/she receives 1 euro, and the participant has to decide whether to keep it (by pressing the *0* on the keyboard) or share it with their partner (by pressing *1* on the keyboard). Deciding to keep the money would yield no earnings for the partner and would end the trial. If participants decide to share, it would result in 5 euros given to the *trustee*, who, in turn, would decide whether: (a) to reciprocate the cooperation, and each of them would receive 2.5€, or (b) not to reciprocate, and the *trustee* would keep the 5€ but the participant would receive nothing. This feedback about the *trustee* decision was displayed on the computer screen 500 ms after the participant took the decision and the trial ended after the feedback. The participants' goal was to maximize their payoffs in the game.

Participants were also informed that they were not playing with real people but that the reciprocity behavior would mimic common patterns of play by real people. Participants were not informed about the different manipulations included in the design: the ethnicity of the interaction partners or the partner's reciprocation rate. Therefore they were unaware of the main goal of the study, which was to explore how the ethnicity of the partners can influence strategies of cooperation, and to investigate whether participants categorize vs. individuate the outgroup vs. the ingroup trustees.

The general procedure was similar to that used by Tortosa et al. (2013). The task was presented on a PC running E-prime software (Schneider et al., 2002). Stimuli were frontal photographs of eight black people (four men and four women) and eight white people (four men–four women) from Nimstim face stimulus set (Tottenham et al., 2009) that represented the *trustees*. Faces were matched on attractiveness and trustworthiness as reported by 28 independent participants (10 men and 18 women, all whites; mean aged 32.68, SD = 6.56) in an online questionnaire using Qualtrics®[2]. All stimuli were presented against a gray background

(see **Figure 1**). Each trial started with a 200 ms presentation of "€" (2.1 × 1.6° visual angle) to indicate the money given to the participant, that was replaced by a fixation point (+, 0.7 × 0.7° visual angle) for 500 ms, and was followed by the picture of the *trustee* for that trial (6.2 × 8.3° visual angle) for 1500 ms. During this time, participants had to indicate whether to keep (by pressing the "0" key) or share (by pressing the "1" key) the euro. After participants informed of their decision (or after 1500 ms in case they did not so), the picture was replaced by the fixation point for 500 ms and then replaced by a symbolic feedback symbol (1.0° × 1.0° visual angle) which indicated the *trustee*'s decision for that trial. Three possible symbols displayed in three different colors were used as feedback: a green "o", a navy "#", and a maroon "*". Their meanings were: "You have decided to keep the money. You receive 1 euro. Your partner receives 0 euro"; "You have decided to share and your partner has decided to reciprocate"; "You have decided to share and your partner has decided not to reciprocate." The association between specific symbols, color, and their meaning was counterbalanced across participants[3]. On trials where participants did not enter their decision on time (1.5 s), they saw the message "¡tarde!" (late!). At the end of the trial a larger fixation point (a "+" sign, 1.0° × 1.0°) remained on the screen for 1000 ms.

Participant played a multi-round design, with 16 different *trustees* over the course of the task. Participants played this game 40 times with each of the 16 *trustees* (for a total of 640 trials) divided in two phases of five blocks each. Each phase was designed so that three faces of an ethnic group were associated with a high probability of reciprocation rate (75%) while three faces of the other group were associated with a low probability of reciprocation rate (25%). These were consistent faces. The forth face of each group in each phase reciprocated at the rate of the other group. These were inconsistent faces. In the second phase the group reciprocation rate was inversed using four different faces for each ethnic group. The order in which black or white started reciprocating in 75% of the trials was counterbalanced across participants. Which face of the group acted as inconsistent face was also counterbalanced across participants. For instance, for a given participant, five blocks constituted the first phase. In

---

[2]There was neither a significant difference in attractiveness between the two ethnic faces $t(27) = 1.64$, $p = 0.11$; $Alpha_{foralltheblackfaces\_attractiveness} = 0.79$, mean = 2.67, SD = 0.61; $Alpha_{forallwhitefaces\_attractiveness} = 0.81$; mean = 2.51, SD = 0.57; nor a significant difference in trustworthiness $t(27) = 0.19$, $p = 0.85$; $Alpha_{foralltheblackfaces\_trustworthiness} = 0.77$, mean = 2.93, SD = 0.52; $Alpha_{forallwhitefaces\_trustworthiness} = 0.71$; mean = 2.91, SD = 0.50.

---

[3]This procedure was used because in a follow up experiment we wanted to evaluate evocated potentials associated to the feedback as have been previously studied in Tortosa et al. (2013).

block 1 reciprocation rate was allocated at 50% for every face of both groups, whereas in blocks 2–5 the reciprocation rate was set at 75% for three black trustees (consistent faces), and 25% for one black trustee (inconsistent face), and 25% for three white trustees (consistent faces) and 75% for one white trustee (inconsistent face). In a second phase of five extra blocks, eight new faces were presented and the reciprocation rates were inversed for the ethnic groups. That is, in the sixth block the reciprocation rate would be again set at 50% for both groups, but in blocks 7–10, three white faces would reciprocate at a 75% rate while three black faces would reciprocate at a 25% rate and one white face would reciprocate at a 25% rate while one black face would reciprocate at a 75% rate.

Once the participants finished the trust game task they were presented with the 16 faces and were asked to evaluate them using a likert-scale ranging from 1 "not at all" to 7 "very much" in what extent they were attractive and trustworthy. We also asked participants to indicate how distinctive the face was compared to the other members of its group, using a likert-scale ranging from −3 (very distinctive) to +3 (very undistinctive), and how frequently each face was presented compared to the others (1 "less," 2 "the same," 3 "more"). We also included some general questions about the group level, including % of reciprocation and % presentation of whites and blacks.

RESULTS

We analyzed the proportion of participants' sharing/cooperation rates across conditions. First, we compared cooperation rate toward black and white trustees in the first block of the first phase (where there was no manipulation of group reciprocation rate, 50%). There was no significant difference in participants cooperation with black (mean = 0.68; SD = 0.16) vs. white trustees (mean = 0.63; SD = 0.19), $t(25) = 1.45$, $p = 0.16$.

In order to measure the categorization or individuation strategies in participants' cooperation behavior, we analyzed it separately for each ethnic group and faces' level of consistency (consistent or inconsistent with their respective category). Thus, cooperation rates were introduced into an ANOVA with ethnicity (black, white), block (2–5), group reciprocation rate (25%, 75%), and face consistency (consistent, inconsistent) as within-subject factors. Result showed that participants, contrary to the social categorization hypothesis, decided to cooperate equally independently of the trustee's ethnicity, $F(1,25) = 0.69$, $p = 0.41$, $\eta_p^2 = 0.03$, that is, they did not cooperate with white trustees more than with black trustees. The main effect of Face consistency was neither significant, $F(1,25) = 0.02$, $p = 0.90$, $\eta_p^2 = 0.00$.

However, and according to our predictions, participants significantly preferred to cooperate with the group associated to high reciprocity ($M = 66.7\%$, CI: 61.4–72.0) as compared to the one associated to low reciprocity ($M = 61.9\%$, CI: 56.0–67.9), $F(1,25) = 13.39$, $p < 0.001$, $\eta_p^2 = 0.35$. This effect of Group Reciprocation rate was significantly moderated by block and Face Consistency, as shown by the three-way interaction between these three factors, $F(1,25) = 4.23$, $p = 0.008$, $\eta_p^2 = 0.15$. The interaction showed that the effect of group reciprocation rate (which was opposite for inconsistent faces) increased across blocks, as learning increased. This makes evident the reinforcement learning

hypothesis (Chang et al., 2010) by which participants update their previous impressions with the acquired knowledge of reciprocity rate of each face.

More interestingly, the Ethnicity by Group reciprocation rate interaction was significant, $F(1,25) = 5.17$, $p = 0.03$, $\eta_p^2 = 0.17$, and was significantly moderated by the Ethnicity × Group Reciprocation × Face consistency three-way interaction, which was also significant, $F(1,25) = 10.47$, $p = 0.003$, $\eta_p^2 = 0.30$. Importantly, contrary to our predictions, a significant Group Reciprocation rate by Face Consistency interaction was observed for black trustees, $F(1,25) = 6.92$, $p = 0.01$, $\eta_p^2 = 0.22$, while the same interaction was not significant for white trustees, $F(1,25) = 0.70$, $p = 0.41$, $\eta_p^2 = 0.03$ (see **Figure 2**). That is, while black trustees led to cooperation responses as a function of the faces' individual cooperation rates (as they were opposite for inconsistent faces), in the case of the white trustees the participant's cooperation behavior was guided by the cooperation rate of the group, independently of the individual cooperation rate (i.e., independently of face consistency). As the same cooperation responses for consistent and inconsistent faces can be conceived as a sign of *categorization*, and opposite cooperation behaviors for consistent vs. inconsistent faces as a sign of *individuation*, these interactions indicated that black trustees were individuated whereas white trustees were categorized.

Trustees' Evaluations

We checked for individual differences of the faces. Specifically, first we wanted to evaluate how the trust game task could have affected judgments of attractiveness and trustworthiness of the trustees. We then performed a repeated measure analysis (two group reciprocation rate by two ethnicity by two face consistency) on each dependent variable. We did not find any significant effect nor interaction for attractiveness, $Fs(1,25) < 2.8$, $ps > 0.11$. Trustworthiness ratings only revealed a significant interaction effect for Ethnicity by Face Consistency, $F(25) = 4.57$, $p = 0.04$, $\eta_p^2 = 0.16$, indicating that consistent black trustees were evaluated as more trustworthy (mean = 3.9; SD = 1.61) than inconsistent black trustees (mean = 3.5; SD = 1.53). However, inconsistent white trustees were evaluated as more trustworthy (mean = 3.7; SD = 1.10) than consistent white trustees (mean = 3.4; SD = 1.32). All other $Fs(1,25) < 2.5$, $ps > 0.14$.

Then we analyzed how distinctive (very distinctive −3 to very undistinctive +3) each individual faces was in comparison with the ingroup faces. The only significant effect was the Group Reciprocation rate by Ethnicity interaction, $F(25) = 4.24$, $p = 0.05$, $\eta_p^2 = 0.15$. The result showed that black trustee associated with low reciprocation rates were perceived as more similar to each other (mean = 1.07; SD = 1.47) than those associated to high reciprocation rates (mean = 0.89; SD = 1.41) while white trustees were perceived as more similar to each other when associated to high group reciprocation rates (mean = 0.96; SD = 1.46) compared to low group reciprocation rates (mean = 0.67; SD = 1.62). None of the other effects reached significance, $Fs(1,25) < 1$, $ps > 0.35$. Next we perform the same analysis to evaluate the perception of the frequency of individual faces presentation (1 "less," 2 "the same," 3 "more" compared to the rest). None of the effects were significant, $Fs(1,25) < 1.5$, $ps > 0.22$.
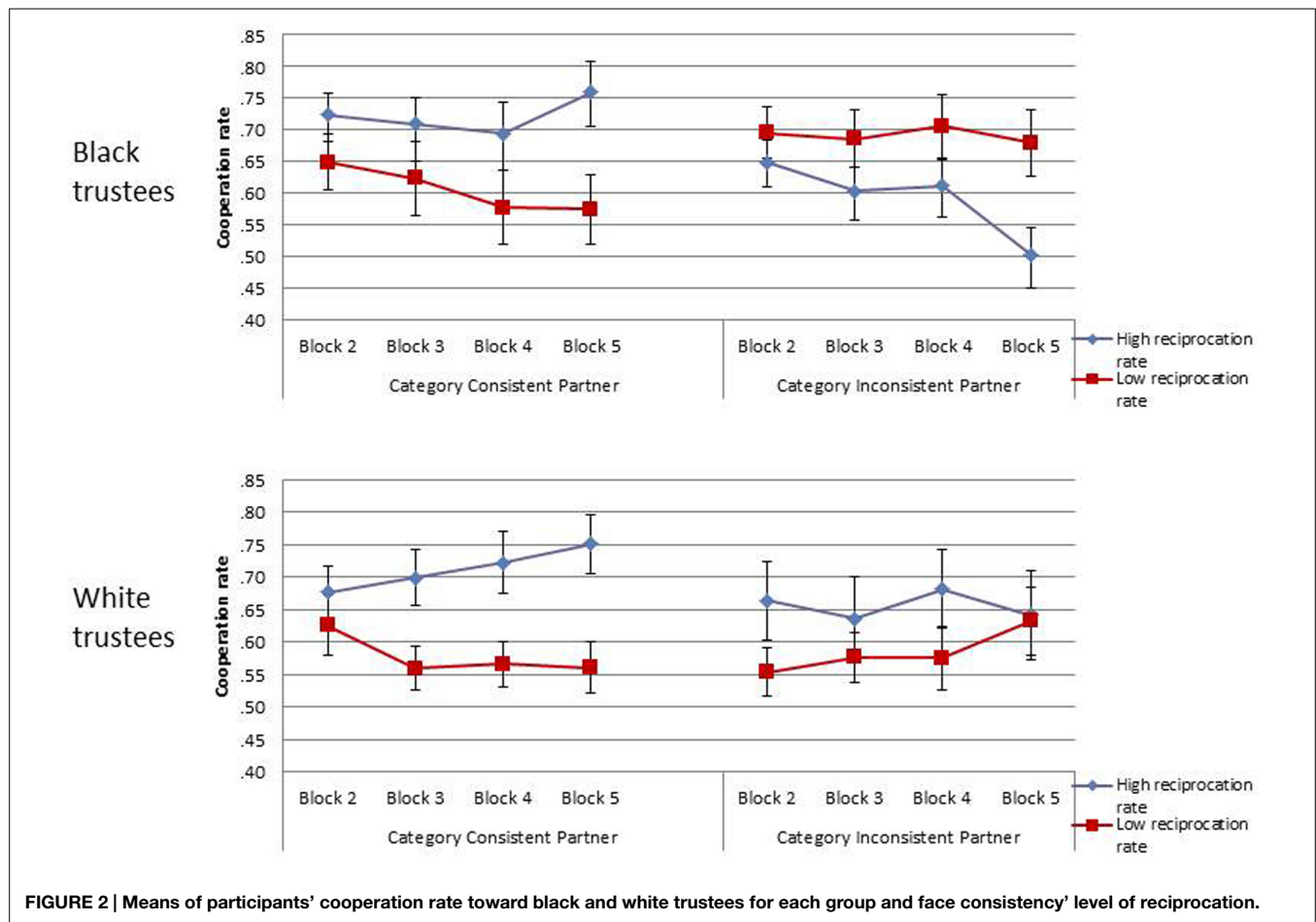
FIGURE 2 | Means of participants' cooperation rate toward black and white trustees for each group and face consistency' level of reciprocation.

We next evaluated how participants perceived the faces at a group level. We first analyzed how frequently participant believed that the two groups of faces (blacks and whites) were presented during the task. There was no significant differences in their estimates of the overall presentation rates of black trustees (mean = 56.15%; SD = 12.11%) compared to white trustees (mean = 51.92%; SD = 15.43), $t(25) = 1.10$, $p = 0.28$, $\eta_p^2 = 0.01$. This result indicates that participants correctly estimated that all faces were equally presented throughout the experiment. Then we evaluated participants' impression about the reciprocation rates of black and white trustees. Interestingly, we found a significant difference in overall reciprocation' judgments depending on the ethnicity of the trustee, $t(25) = 2.95$, $p = 0.005$, $\eta_p^2 = 0.26$. Participants reported they thought that black trustees reciprocated more often (mean = 60.58%; SD = 13.44) than white trustees (mean = 48.50%; SD = 16.00).

DISCUSSION

The present study explored the effect of ethnicity and consistent vs. inconsistent behaviors (reciprocation rates) regarding their identity group in a multi-round trust game task. We wanted to explore whether ethnicity moderates the decision of whether to cooperate with partners or not and, more importantly, whether social categorization or individuation processes would underlie

those decisions. Results revealed that participants did not show a particular bias toward cooperating with white compared to black in general, although, interestingly, they used different strategies to make decisions about how to cooperate (share money) or not with white and black partners. Whereas the observed pattern of results led us to conclude that the white ingroup trustee's faces were categorized (i.e., the same cooperation pattern was observed for consistent and inconsistent faces), the black outgroup trustee's faces were individuated (i.e., an opposite pattern of cooperation was observed for consistent and inconsistent faces).

Even though preferences to cooperate with ingroup members more than with outgroup members have been largely reported in previous research (Wilson and Kayatani, 1968; Tanis and Postmes, 2005; Chang et al., 2010), other studies' results go in opposite direction, that is, favoring outgroup members (see Allport, 1954; Monteith et al., 2002; Tortosa et al., 2013, study 2). However, our results did not show any bias neither for black nor for white trustees as measured in block 1 (50% reciprocation rate for both black and white trustees). This finding is in line with previous results by Stanley et al. (2011) which show that unless participants had a strong pro-whites or pro-black bias, as measure with an implicit ethnic attitudes test, their evaluation of trustworthiness and their cooperation behavior (economic offers in a trust game) kept similar toward black and white partners.

We can rule out the possibility that black and white trustees evoked different trustworthiness impression, as we controlled for this (among other variables, e.g., attractiveness) with the pretest for stimuli selection. The evaluation of the trustworthiness of the stimuli at the end of the trust game did not show either overall differences between black and white trustees, which go in line with the pretest and with other studies investigating ethnic attitudes (Phelps et al., 2000; Stanley et al., 2011).

A potential explanation for the similar cooperation toward partners belonging to both ethnic groups can be due to the use of women and men as stimuli. It may exist a confound between these two groups, so participants prefer to cooperate with women more as they are perceived more trustworthy than men (independently of their ethnic categorization, Buchan et al., 2008) and consequently the gender bias may have concealed the ethnic bias. This is surely a confound factor that should be carefully analyzed by future research.

Interestingly, the manipulation of consistency significantly affected the evaluation of trustworthiness, which may explain the current results in our study. The different evaluation of inconsistent black and white faces being the former more positively evaluated regarding trustworthiness than the later may evidence that people accepted more ingroup members (whites) that behave unexpectedly compared to outgroup members (blacks; Kosic et al., 2014).

Our main contribution to the study of ethnic categories and decision-making literature focuses on the study of cooperation strategies related to categorization and individuation processes. Result showed that (white) participants used different strategies to make decisions on how to cooperate (share money) with white and black partners. Specially, they learnt which face is behaving inconsistently with the rest of the group and decided how to cooperate with this person accordingly to the specific cooperation rate that he or she showed. That is, participants individuate each trustee they were encountering with. Interestingly, however, this individuation strategy applied exclusively to black faces (outgroup members). Contrary, decisions to cooperate toward white trustees followed a categorization strategy. That is, participants took their decisions to cooperate with inconsistent trustees as a function of the proportion of reciprocation assigned to the majority of the white trustees (consistent trustees).

The reason why participants categorize whites and individuate blacks, contrary to our expectations, and to what was previously shown (Hugenberg et al., 2010) is far from being clear. However, it could be argued that participants may care about ingroup identification (Castano and Yzerbyt, 1998); therefore, they may be motivated to preserve the homogeneity of the ingroup members (Castano, 1999) producing the categorization effect observed for white faces. Furthermore, according to interdependence theories, participants may have individuated black faces given that their outcomes (the money they could earn during the task) may depend on their sharing behaviors (Ruscher and Fiske, 1990). Participants may have paid special attention to black people to compensate their dispositional behavior to categorize them and by consequence they increased their attention to inconsistencies among black partners. This increased attention may have helped

them to use the strategy to cooperate with each face according to the individual reciprocation rate rather than the group reciprocation rate.

Another alternative explanation comes from Collins et al. (2011) model of learning phenomena, and concretely the "blocking" (Kamin, 1968) explanation, explaining why people learnt with different strategies about black and white partners. "Blocking" might occurs for whites when a new proportion of reciprocity (inconsistent-cue) is introduced alongside a proportion of reciprocity (consistent-cue) whose meaning has already been learned about the majority of the members of the group. Because the perceptual information coming from the inconsistent partner (white person) is redundant at the perceptual level (providing no additional information beyond the original cue), learning about it may have been blocked.

Interestingly, while blocking could explain the null effect (more related to categorization for whites), highlighting could explain the individuation effect for blacks. Highlighting occurs when a person focuses extra attention on a cue that changes the meaning of a previously learned cue, as happens when a learned association is no longer correct when a new cue is added alongside a known one (Kruschke, 2009). Another explanation to blocking from a motivational perspective will indicate that for white participants it is not enriching on a matter of novelty to learn about others whites, but it is highly interesting to know about the outgroup, to avoid threats (highlighting).

Unfortunately we do not have information about participants' previous experience with black individuals, so future studies should measure and control for it. Future research should also focus in explaining the mechanism underlying the individuation—categorization strategies chosen by the participants, not only in economic games, but also in other social interactions, such as prosocial behaviors. It will be also interesting to know whether bottom-up (perceptual information) or top-down (conceptual-stereotypes) processes influence judgmental tasks. Previous research in gender-emotion stereotypes (Becker et al., 2007) show that both top-down and bottom-up processes can co-occur during people evaluation.

Another specific detail of our procedure is that another group category apart from the ethnical group (i.e., gender) could be salient, as half of the faces in each group were women whereas the other half were men. Given that the majority of participants in our study (all but one) were women this might have affected the pattern of results. However, given that this occurred for the ethnic groups it seems unlikely that it could explain the pattern of results. Nevertheless, future research should control more carefully for the presence/absence of different important category features (ethnicity, gender, age, etc.).

It seems clear that future research is necessary to replicate and consolidate the specific findings observed in the reported study, and to better explain the observed pattern of results. Nevertheless, and importantly, the current study has shown to be a suitable tool to investigate the incidental generation and use of categorization-individuation social cooperation processes. In a previous study (Canadas et al., 2013), this general paradigm showed to be also suitable to investigate these categorization-individuation processes and their use underlying the implicit

allocation of attentional control. We believe this paradigm could be extended to the study of other situations where categorization vs. individuation processes play an important role in social interactions. Perhaps the individuation pattern observed for the outgroup members might disappear whenever more than four members from each category have to be tracked. Therefore, our procedure might be useful to investigate the interplay between using specific knowledge about our interaction with a particular individual to predict his/her future behavior vs. using knowledge

we have about our previous interactions with other members of the same group, and the boundary conditions for the use of one process or the other.

ACKNOWLEDGMENTS

REFERENCES

Allport, G. W. (1954). *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley.

Axelrod, R., and Hamilton, W. D. (1981). The evolution of cooperation. *Science* 211, 1390–1396. doi: 10.1126/science.7466396

Balliet, D., and Van Lange, P. A. M. (2013). Trust, punishment and cooperation across 18 societies: a meta-analysis. *Perspect. Psychol. Sci.* 8, 363–379. doi: 10.1177/1745691613488533

Bargh, J. A., and Williams, E. L. (2006). The automaticity of social life. *Curr. Dir. Psychol. Sci.* 15, 1–4. doi: 10.1111/j.0963-7214.2006.00395.x

Becker, D. V., Kenrick, D. T., Neuberg, S. L., Blackwell, K. C., and Smith, D. M. (2007). The confounded nature of angry men and happy women. *J. Pers. Soc. Psychol.* 92, 179–190. doi 10.1037/0022-3514.92.2.179

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games Econ. Behav.* 10, 122–142. doi: 10.1006/game.1995.1027

Bolle, F. (1998). Rewarding trust: an experimental study. *Theory Decis.* 45, 83–98. doi: 10.1023/A:1004930101743

Brewer, M. B. (1981). "Ethnocentrism and its role in interpersonal trust," in *Scientific Inquiry and the Social Sciences*, eds M. B. Brewer and B. E. Collins (San Francisco: Jossey-Bass), 345–359.

Brewer, M. B. (1988). "A dual process model of impression formation," in *Advances in Social Cognition*, Vol. 1, eds T. K. Srull and J. R. S. Wyer (Hillsdale, NJ: Erlbaum), 1–36.

Buchan, N. R., Croson, R. T. A., and Solnick, S. (2008). Trust and gender: an examination of behavior and beliefs in the Investment Game. *J. Econ. Behav. Organ.* 68, 466–476. doi: 10.1016/j.jebo.2007.10.006

Camerer, C., and Weigelt, K. (1988). Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56, 1–36. doi: 10.2307/1911840

Campellone, T. R., and Kring, A. M. (2013). Who do you trust? The impact of facial emotion and behaviour on decision making. *Cogn. Emot.* 27, 603–620. doi: 10.1080/02699931.2012.726608

Canadas, E., Rodriguez-Bailon, R., Milliken, B., and Lupianez, J. (2013). Social categories as a context for the allocation of attentional control. *J. Exp. Psychol. Gen.* 142, 934–943. doi: 10.1037/a0029794

Castano, E. (1999). *The Phenomenology of the Ingroup: Entitativity and Identification*. Ph.D. thesis, Catholic University of Louvain, Louvain-la-Neuve.

Castano, E., and Yzerbyt, V. Y. (1998). The highs and lows of group homogeneity. *Behav. Process.* 42, 219–238. doi: 10.1016/s0376-6357(97)00078-8

Castelli, L., Zogmaister, C., Smith, E. R., and Arcuri, L. (2004). On the automatic evaluation of social exemplars. *J. Pers. Soc. Psychol.* 86, 373–387. doi: 10.1037/0022-3514.86.3.373

Chang, L. J., Doll, B. B., Van 't Wout, M., Frank, M. J., and Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cogn. Psychol.* 61, 87–105. doi: 10.1016/j.cogpsych.2010.03.001

Charness, G., and Rabin, M. (2002). Understanding social preferences with simple tests. *Q. J. Econ.* 117, 817–869. doi: 10.1162/003355302760193904

Chen, B., Zhang, B., and Zhu, W. (2010). Combined trust model based on evidence theory in iterated prisoner's dilemma game. *Int. J. Syst. Sci.* 42, 63–80. doi: 10.1080/00207720903434771

Collins, E. C., Percy, E. J., Smith, E. R., and Kruschke, J. K. (2011). Integrating advice and experience: learning and decision making with social and nonsocial cues. *J. Pers. Soc. Psychol.* 100, 967–982. doi: 10.1037/a0022982

Cuddy, A. J. C., Fiske, S. T., and Glick, P. (2004). When professionals become mothers, warmth doesn't cut the ice. *J. Soc. Issues* 60, 701–718. doi: 10.1111/j.0022-4537.2004.00381.x

Cunningham, W. A., and Zelazo, P. D. (2007). Attitudes and evaluations: a social cognitive neuroscience perspective. *Trends Cogn. Sci.* 11, 97–104. doi: 10.1016/j.tics.2006.12.005

Devine, P. G. (1989). Stereotypes and prejudice: their automatic and controlled components. *J. Pers. Soc. Psychol.* 56, 5–18.

Dunning, D., Anderson, J. E., Schlosser, T., Ehlebracht, D., and Fetchenhauer, D. (2014). Trust at zero acquaintance: more a matter of respect than expectation of reward. *J. Pers. Soc. Psychol.* 107, 122–141. doi: 10.1037/a0036673

Fehr, E., and Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends Cogn. Sci.* 11, 419–427. doi: 10.1016/j.tics.2007.09.002

Fehr, E., and Fischbacher, U. (2002). Why social preferences matter—the impact of non-selfish motives on competition, cooperation and incentives. *Econ. J.* 112, C1–C33. doi: 10.1111/1468-0297.00027

Fehr, E., and Fischbacher, U. (2003). The nature of human altruism. *Nature* 425, 785–791. doi: 10.1038/nature02043

Fehr, E., and Schmidt, M. (1999). A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868. doi: 10.1162/003355399556151

Fiske, S. T. (2003). "Five core social motives, plus or minus five," in *Motivated Social Perception: The Ontario Symposium*, Vol. 9, eds S. J. Spencer, S. Fein, M. Zanna, and J. Olson (Mahwah, NJ: Erlbaum), 233–246.

Fiske, S. T., and Neuberg, S. L. (1990). "A continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation," in *Advances in Experimental Social Psychology*, Vol. 23, ed. M. P. Zanna (San Diego, CA: Academic Press), 1–108.

Frank, R. H., Gilovich, T., and Regan, D. T. (1993). The evolution of one-shot cooperation: an experiment. *Ethol. Sociobiol.* 4, 247–256. doi: 10.1016/0162-3095(93)90020-I

Gaertig, C., Moser, A., Alguacil, S., and Ruz, M. (2012). Social information and economic decision-making in the ultimatum game. *Front. Neurosci.* 6:103. doi: 10.3389/fnins.2012.00103

Hobbes, T. (1997). *The Leviathan*. New York, NY: Touchstone Press. (Original work published 1660).

Hugenberg, K., Young, S. G., Bernstein, M. J., and Sacco, D. F. (2010). The categorization-individuation model: an integrative account of the other-race recognition deficit. *Psychol. Rev.* 117, 1168–1187. doi: 10.1037/a0020463

Johnson, N. D., and Mislin, A. A. (2011). Trust games: a meta-analysis. *J. Econ. Psychol.* 32, 865–889. doi: 10.1016/j.joep.2011.05.007

Judd, C. M., and Park, B. (1988). Out-group homogeneity: judgments of variability at the individual and group levels. *J. Pers. Soc. Psychol.* 54, 778–788. doi: 10.1037/0022-3514.54.5.778

Kamin, L. J. (1968). ""Attention-like" processes in classical conditioning," in *Miami Symposium on the Prediction of Behavior, 1967: Aversive Stimulation*, ed. M. R. Jones (Coral Gables, Florida: University of Miami Press), 9–31.

Kawakami, K., Dion, K. L., and Dovidio, J. F. (1998). Racial prejudice and stereotype activation. *Pers. Soc. Psychol. Bull.* 24, 407–416. doi: 10.1177/0146167298244007

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., and Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78–83. doi: 10.1126/science.1108062

Kosic, A., Mannetti, L., and Livi, S. (2014). Forming impressions of in-group and out-group members under self-esteem threat: the moderating role of the need for cognitive closure and prejudice. *Int. J. Intercult. Relat.* 40, 1–10. doi: 10.1016/j.ijintrel.2014.01.004

Kruschke, J. K. (2009). "Highlighting: a canonical experiment," in *The Psychology of Learning and Motivation*, Vol. 51, ed. B. Ross (San Diego: Academic Press), 153–185.

Lepore, L., and Brown, R. (1997). Category and stereotype activation: is prejudice inevitable? *J. Pers. Soc. Psychol.* 72, 275–287. doi: 10.1037/0022-3514.72.2.275

Levin, D. T. (1996). Classifying faces by race: the structure of face categories. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1364–1382. doi: 10.1037/0278-7393.22.6.1364

Levin, D. T. (2000). Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the CR recognition deficit. *J. Exp. Psychol. Gen.* 129, 559–574. doi: 10.1037/0096-3445.129.4.559

Linville, P. W., Fischer, G. W., and Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. *J. Pers. Soc. Psychol.* 57, 165–188. doi: 10.1037/0022-3514.57.2.165

Machiavelli, N. (1515/2003). *The Prince*, trans. R. Goodwin. Wellesley, MA: Dante University Press.

MacLin, O. H., and Malpass, R. S. (2001). Racial categorization of faces: the ambiguous race face effect. *Psychol. Public Policy Law* 7, 98–118. doi: 10.1037/1076-8971.7.1.98

Macrae, C. N., and Cloutier, J. (2009). A matter of design: priming context and person perception. *J. Exp. Soc. Psychol.* 45, 1012–1015. doi: 10.1016/j.jesp.2009.04.021

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., and Czopp, A. M. (2002). Putting the brakes on prejudice: on the development and operation of cues for control. *J. Pers. Soc. Psychol.* 83, 1029–1050. doi: 10.1037/0022-3514.83.5.1029

Nelson, T. D. (2005). Ageism: prejudice against our feared future self. *J. Soc. Issues* 61, 207–221. doi: 10.1111/j.1540-4560.2005.00402.x

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *J. Cogn. Neurosci.* 12, 729–738. doi: 10.1162/089892900562552

Platow, M. J., Foddy, M., Yamagishi, T., Lim, L., and Chow, A. (2012). Two experimental tests of trust in in-group strangers: the moderating role of common knowledge of group membership. *Eur. J. Soc. Psychol.* 42, 30–35. doi: 10.1002/ejsp.852

Ruscher, J. B., and Fiske, S. T. (1990). Interpersonal competition can cause individuating processes. *J. Pers. Soc. Psychol.* 58, 832–843. doi: 10.1037/0022-3514.58.5.832

Ruz, M., Moser, A., and Webster, K. (2011). Social expectations bias decision-making in uncertain inter-personal situations. *PLoS ONE* 6:e15762. doi: 10.1371/journal.pone.0015762

Ruz, M., and Tudela, P. (2011). Emotional conflict in interpersonal interactions. *Neuroimage* 54, 1685–1691. doi: 10.1016/j.neuroimage.2010.08.039

Scharlemann, J. P. W., Eckel, C. C., Kacelnik, A., and Wilson, R. K. (2001). The value of a smile: game theory with a human face. *J. Econ. Psychol.* 22, 617–640. doi: 10.1016/S0167-4870(01)00059-9

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-Prime User's Guide*. Pittsburgh, PA: Psychology Software Tools Inc.

Solnick, S. J. (2001). Gender differences in the ultimatum game. *Econ. Inq.* 39, 189–200. doi: 10.1093/ei/39.2.189

Solnick, S. J., and Schweitzer, M. (1999). The influence of physical attractiveness and gender on ultimatum game decisions. *Organ. Behav. Hum. Decis. Process.* 79, 199–215. doi: 10.1006/obhd.1999.2843

Sommers, S. R. (2006). On racial diversity and group decision making: identifying multiple effects of racial composition on jury deliberations. *J. Pers. Soc. Psychol.* 90, 597–612. doi: 10.1037/0022-3514.90.4.597

Sommers, S. R. (2007). Race and the decision making of juries. *Legal Criminol. Psychol.* 12, 171–187. doi: 10.1348/135532507X189687

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., and Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7710–7715. doi: 10.1073/pnas.1014345108

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

Tanis, M., and Postmes, T. (2005). A social identity approach to trust: interpersonal perception, group membership and trusting behaviour. *Eur. J. Soc. Psychol.* 35, 413–424. doi: 10.1002/ejsp.256

Tortosa, M. I., Lupianez, J., and Ruz, M. (2013). Race, emotion and trust: an ERP study. *Brain Res.* 1494, 44–55. doi: 10.1016/j.brainres.2012.11.037

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., et al. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Res.* 168, 242–249. doi: 10.1016/j.psychres.2008.05.006

Wilson, W., and Kayatani, N. (1968). Intergroup attitudes and strategies in game between opponents of the same or of a different race. *J. Pers. Soc. Psychol.* 9, 24–30. doi: 10.1037/h0025720