Implicit race attitudes predict trustworthiness judgments and economic trust decisions

Damian A. Stanley[a], Peter Sokol-Hessner[a], Mahzarin R. Banaji[b], and Elizabeth A. Phelps[a,1]

[a]Department of Psychology, New York University, New York, NY 10003; and [b]Department of Psychology, Harvard University, Cambridge, MA 02138

Edited by Daniel Kahneman, Princeton University, Princeton, NJ, and approved March 23, 2011 (received for review September 23, 2010)

Trust lies at the heart of every social interaction. Each day we face decisions in which we must accurately assess another individual's trustworthiness or risk suffering very real consequences. In a global marketplace of increasing heterogeneity with respect to nationality, race, and multiple other social categories, it is of great value to understand how implicitly held attitudes about group membership may support or undermine social trust and thereby implicitly shape the decisions we make. Recent behavioral and neuroimaging work suggests that a common mechanism may underlie the expression of implicit race bias and evaluations of trustworthiness, although no direct evidence of a connection exists. In two behavioral studies, we investigated the relationship between implicit race attitude (as measured by the Implicit Association Test) and social trust. We demonstrate that race disparity in both an individual's explicit evaluations of trustworthiness and, more crucially, his or her economic decisions to trust is predicted by that person's bias in implicit race attitude. Importantly, this relationship is robust and is independent of the individual's bias in explicit race attitude. These data demonstrate that the extent to which an individual invests in and trusts others with different racial backgrounds is related to the magnitude of that individual's implicit race bias. The core dimension of social trust can be shaped, to some degree, by attitudes that reside outside conscious awareness and intention.

decision making | implicit bias | social attitudes | behavioral economics | Trust Game

Social trust is critical for the decisions and actions that underlie the smooth functioning of any society (1). We routinely decide whether to trust and in whom to trust in social situations, thereby exposing ourselves to the risk of loss for the possibility of greater reward. Such is the case in a broad range of social interactions, from the interpersonal (trusting a confidant) to the economic (trusting a financial advisor) to the political (trusting a candidate).

In many instances, we have prior experience with potential partners who may be family members, friends, or business associates and can rely on that experience when evaluating their trustworthiness. Absent such experience, we use available information that our culture and experiences signal to be diagnostic of trust, such as the other person's social group. Recent psychological and neuroimaging work has demonstrated that such trustworthiness evaluations are made rapidly (<100 ms) (2) and automatically (3–5). These properties confer a clear benefit on the trustor when a given social preconception accurately predicts the trustee's behavior.

In the complex demographic milieu of modern society, such preconceptions can also lead to faulty evaluations of trustworthiness. In a recent but particularly infamous example, financier Bernard Madoff's own group membership played a role in eliciting trust from a large number of Jewish investors who clearly overestimated his trustworthiness, in part, because of a shared group identity (6). Although the Madoff case represents an extreme example of trust based on group membership gone awry, it is conceivable that many ordinary decisions of trust involve similar reliance on the social categories of others, whether they ultimately serve the decision maker well or not.

Although much attention has been paid recently to the question of trust (2–5, 7–10), the majority of studies focus on the common factors that contribute to trust (e.g., situational factors or characteristics of those to be trusted). Decisions in the worlds of business, law, education, and medicine, and even more ordinary daily interactions between individuals, all rely on trust. Increasingly, that trust must be forged between individuals who differ in background, shared experiences, and aspirations. We currently know almost nothing about the role of individual differences in estimating trustworthiness, or how these individual differences on the part of the person deciding to trust may interact with the social group of the person being trusted. In other words, what may lead one person to be trusted by some and distrusted by others. In this research, we explore the degree to which our individual implicit associations concerning social groups dictate in whom we trust.

There is now common agreement that a useful distinction exists between explicit and implicit mental processes, including attitudes, beliefs, and self-perceptions (11). Implicit mental processes, expected to operate relatively automatically and without awareness, can also be oppositional to our intended goals (12, 13). Here, we focus on implicit social bias, a measure of how strongly one associates a concept (e.g., pleasant/unpleasant) with one or another social group. Recent work on measures of implicit biases has shown that they are pervasive and robust (14) and that they can predict social behaviors (15, 16), including the decisions of highly trained professionals, such as doctors (12, 17).

The psychological and neural mechanisms subserving evaluations of trustworthiness and certain types of implicit social biases may overlap. The rapid (2) and implicit (3–5) nature of trustworthiness evaluations suggests a reliance on automatic processes, such as previously stored social preferences, which do not require conscious reflection to be expressed. Supporting this, the amygdala, a subcortical group of nuclei involved in automatic processing of emotional stimuli and fear learning (18), has been implicated in both trustworthiness evaluations (3, 4) and the expression of race-related implicit biases (19–21) [although it is not necessary for the behavioral expression of race-related implicit bias (22)]. Other research has shown that in economic decision making, social information pertaining to a partner's moral character modulates reward-related blood oxygen level-dependent responses, even when participants explicitly know that the partner's moral character does not predict cooperativeness (7). Such evidence of process similarity, anatomical overlap, and reward modulation is suggestive, but it is only that. To date, no studies have demonstrated the existence of a direct link, behavioral or otherwise, between an individual's implicit social bias for some groups over others and his or her trustworthiness evaluations of members of those groups.

Author contributions: D.A.S., P.S.-H., M.R.B., and E.A.P. designed research; D.A.S. performed research; D.A.S., P.S.-H., M.R.B., and E.A.P. analyzed data; and D.A.S., P.S.-H., M.R.B., and E.A.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: liz.phelps@nyu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1014345108/-/DCSupplemental.

Motivated by the evidence for overlap in process and mechanism, we investigated the relationship between an individual's level of implicit race bias and his or her estimations of the trustworthiness of others. To assess implicit race attitude, we used a common version of the Implicit Association Test (IAT) (23) that tests the strength of an individual's automatic association between black and white male faces and the concepts of pleasant and unpleasant (14). The results of IATs like this vary from individual to individual and are assumed to reflect social experience (24). To assess disparity in estimations of trustworthiness, we had participants rate the trustworthiness of others' faces (study 1) and make realistic economic decisions in a series of interpersonal economic games (study 2).

Study 1: Trustworthiness Ratings

We first examined whether an individual's implicit black/white race attitude, as measured by the IAT (23), predicted disparity in his or her trustworthiness evaluations of unfamiliar black and white males. Participants viewed a large set of pictures of emotionally neutral male faces (100 black, 100 white, and 91 other race) and rated their trustworthiness on a scale from 1 (not at all trustworthy) to 9 (extremely trustworthy). Other-race faces were included to provide a more representative set of faces for evaluation so that participants were unaware of a focus on race. Immediately following the ratings task, we assessed participants' implicit race attitude and then their explicit race attitudes. If individual differences in implicit race attitude (pro-black or pro-white) are predictive of trust, we should observe a relationship between the magnitude of implicit race bias and judgments of trustworthiness of targets belonging to these groups. Unlike many studies of race attitudes, we explicitly chose not to restrict participation on the basis of ethnicity or race. This methodological choice was made both because of our focused interest in the psychological construct of implicit attitudes rather than simple group-level effects and to allow us to explore individual differences in those attitudes by ensuring that our sample had a wide range of implicit race attitudes from pro-black to pro-white. Nevertheless, we ensured that our participants' ethnicity could not account for our findings by examining the role of participant race (white/nonwhite) in each of our main results.

We wish to emphasize here that there is not a simple correspondence between individuals' implicit racial attitudes and their own race (14). This is both because the former is a continuous, objective, and quantifiable variable, whereas the latter is categorical (and increasingly subjective), and because implicit attitudes are thought to result from many sources beyond one's own race, including environmental exposure (25) and personal interactions (26). In this series of experiments, we sought to examine the continuous relationship between individuals' implicit racial attitudes and trustworthiness estimations.

Results: Study 1

The mean trustworthiness rating for all faces on a scale of 1–9 was 5.02 (range: 1–9, SD = 0.91, n = 50). Analysis of within-race mean face ratings found no significant difference between black (μ = 4.82, SD = 0.95) and white (μ = 5.07, SD = 1.10) faces, although there was a trend for white faces to be rated as slightly more trustworthy than black faces [$t(49)$ P = 0.073; two-tailed paired Student's t test using the z-scored ratings data].* This trend was also present when considering the responses of white participants alone [$t(30)$ P = 0.088] but not those of nonwhite participants [$t(18)$ P = 0.581], although there was no statistical difference between the two. The mean IAT score was 0.41 (SD = 0.41; μ = 0.44, SD = 0.41 for white participants; μ = 0.36, SD = 0.42 for nonwhite participants; and no significant difference between the two). Group means for the other measures we collected are reported in *SI Text* (Table S1). The lack of strong differences in the perceived trustworthiness of faces from different races is in accord with other studies investigating implicit race attitudes (19, 12). In those studies, the analysis of how individual differences in the dependent variable covaried with measures of implicit race bias was critical for uncovering the relationship between the two.

Our analysis focused on the continuous relationship between individual differences in race attitudes and perceived trustworthiness. We found that differences in implicit race attitudes (IAT D score) predicted disparity in the perceived trustworthiness of black and white faces. Individuals whose IAT scores reflected a stronger pro-white implicit bias were likely to judge white faces as more trustworthy than black faces, and vice versa. We quantified each individual's black/white disparity in perceived trustworthiness by converting all 291 ratings to z scores (analysis of raw responses is provided in *SI Text*; Fig. S1) and then subtracting the mean black score from the mean white score for that individual. Individual differences in IAT score were significantly correlated (Pearson's r) with individual differences in rating disparity [$r(48)$ = 0.4182, P = 0.0025 across all participants, Fig. 1*B*; white participants only, $r(29)$ = 0.52, P = 0.0025; nonwhite participants only, $r(17)$ = 0.15, P = 0.54]. Although the correlation for nonwhite participants was not significant, we note it was consistent in direction with the other correlations. Confirming this finding, a robust regression found that IAT score predicted rating disparity (β_{IAT} = 0.51, P = 0.008), whereas a dummy predictor for white vs. nonwhite participants did not ($\beta_{Participant\ Race}$ = 0.06, P = 0.69). Overall, this analysis demonstrated the existence of

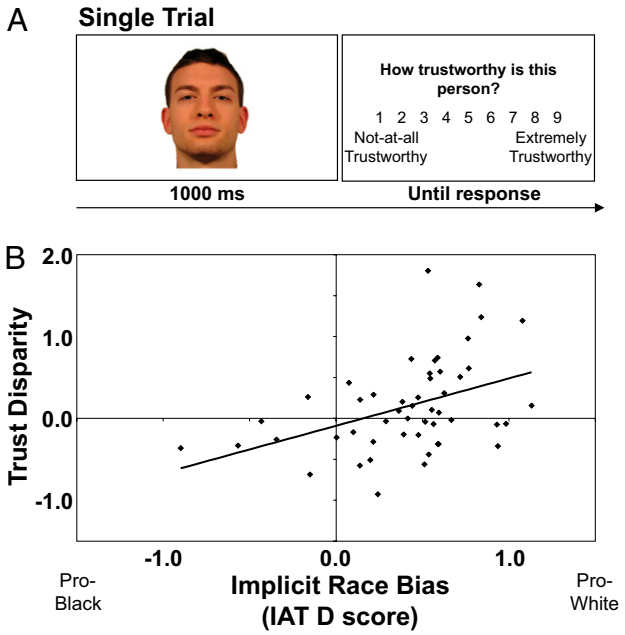

Fig. 1. Study 1: Individual differences in implicit race attitude correlate with race disparity in trustworthiness evaluations. (*A*) Diagram of a single trustworthiness rating trial. (*B*) Scatter plot showing each individual's score on the black/white, pleasant/unpleasant IAT and their black/white disparity in trustworthiness ratings [Pearson's $r(48)$ = 0.4182; P = 0.0025]. Trust Disparity = Mean(white rating z score) − Mean(black rating z score).

*The presence or absence of group-level disparities is not relevant to our central question, namely, whether the degree of implicit bias predicts the degree of trustworthiness rating disparity across individuals. Taking the example of a linear regression, although the constant term (i.e., group-level disparities) can be informative, our interest was in the slope (i.e., the continuous quantitative relationship of implicit bias to estimations of trustworthiness).

a roughly linear relationship between implicit racial attitudes and disparity in ratings of trustworthiness.

We further investigated the strength of the relationship between IAT race attitude scores and rating disparity as well as the independence of that relationship from any analogous association with explicit measures of race attitude or participant race (white or nonwhite). A bootstrap analysis (details are provided in *SI Text*; Fig. S2) found the correlation between IAT scores and rating disparity to be highly robust; individual differences in IAT score were positively correlated with rating disparity in more than 95% of random samples of nine or more ratings (three from each race category). More importantly, the influence of implicit race attitude on evaluations of trustworthiness was independent of that of explicit race attitude. We conducted a stepwise regression analysis of the race disparity in ratings to objectively assess the relative contributions of implicit race attitude (IAT D score), four standard measures of explicit race attitude (27–29), a measure of political leaning (30), and participant race (Table 1). In the final model, IAT score remained a significant independent predictor of race disparity in trustworthiness ratings, accounting for a portion of the variance that standard measures of explicit race attitude, political leaning, and participant race did not. (For both studies 1 and 2, we do not offer an interpretation of the explicit measures that remained significant factors in the final regression model because their inclusion was used primarily to demonstrate the independence of the relationship between implicit race attitudes and trust rather than to investigate the role of explicit attitudes.)

Although these data are quite compelling, the rating task of study 1 lacks ecological validity in a number of ways. Rarely in the course of everyday life do we explicitly evaluate someone's trustworthiness. More often, that evaluation is implicit, inherent to the social interactions in which we find ourselves. In addition, ratings do not capture the context of real-world trustworthiness evaluations (implicit or explicit) that occur while making decisions with potential consequences. To address these issues and extend our findings, we adapted a paradigm from behavioral economics. Specifically, we used a modified version of the Trust Game (8) to characterize the relationship between implicit race attitude and decisions in potentially beneficial but risky economic interactions.

Study 2: Economic Offers

A different group of participants each played a series of single-shot modified Trust Games (8) with 291 distinct partners (the exact photographs used in study 1; Fig. 2*A*). In each interaction, the participant chose how much to offer a partner ($0–$10), with the understanding that the partner would receive quadruple the amount the participant offered. Participants were told that their partners, depicted in the photographs, were real individuals the experimenter had previously interviewed and who had already made the decision to return half or keep all of whatever amount they received. Thus, participants had to judge whether their partner had made a mutually beneficial decision (in which case, the participant could increase their payoff) or if they had acted selfishly. Note that at no point were the participants asked to explicitly evaluate the trustworthiness of their partners. Rather, the measure of trust was an ecologically relevant consequential decision about how much money to risk in each interaction. Following the completion of the Trust Game, we assessed participants' implicit and explicit race attitudes as in study 1 (IAT, followed by explicit measures). Based on the findings of study 1, we hypothesized that the magnitude of participants' bias in implicit race attitude (pro-black or pro-white) would predict an overall disparity in their monetary offers to members of each group. To ensure, again, a range of implicit race attitudes in our sample, and the generality of our findings, participation was not restricted on the basis of race or ethnicity.

Results: Study 2

The pattern of results for monetary offers in study 2 was similar to that of the ratings in study 1. On average, black and white partners were offered similar amounts of money. The overall mean offer ($0–$10) was $3.77 (SD = $1.77, n = 43 of 57; ex-

Table 1. Bias in implicit race attitude predicts trust disparity independent of explicit race attitude

Experiment: Trustworthiness ratings, n = 48			
Dependent variable: Trust disparity			
Final model: r^2 = 0.329, P < 0.001			
Independent predictors	IAT, EMS, IMS, MRS, SRS, LIB/CON, participant race		
Factors	Standardized β	Significance in final model (P)	Change in r^2
IAT	0.376	0.003	0.142
LIB/CON	0.430	<0.001	0.187
Experiment: Modified Trust Game, n = 43			
Dependent variable: Offer disparity Final model: r^2 = 0.247, P < 0.003			
Independent predictors	IAT, EMS, IMS, MRS, SRS, LIB/CON, participant race		
Factors	Standardized β	Significance in final model (P)	Change in r^2
IAT	0.358	0.014	0.166
EMS	0.289	0.044	0.081

Separate stepwise regression analyses (probability of F to enter, P = 0.05; probability of F to remove, P = 0.10) for disparity in ratings (study 1) and offers (study 2) found that IAT scores independently accounted for a significant portion of the variance in both, even when accounting for explicit race attitudes and participant race (n = 48 for the analysis of data from study 1 because two participants did not complete the explicit measures portion of the experiment). EMS, External Motivation to Avoid Prejudice Survey; IMS, Internal Motivation to Avoid Prejudice Survey; MRS, Modern Racism Scale; SRS, Symbolic Racism Scale; LIB/CON, political leaning scale (Liberal/Conservative); participant race, white/nonwhite.

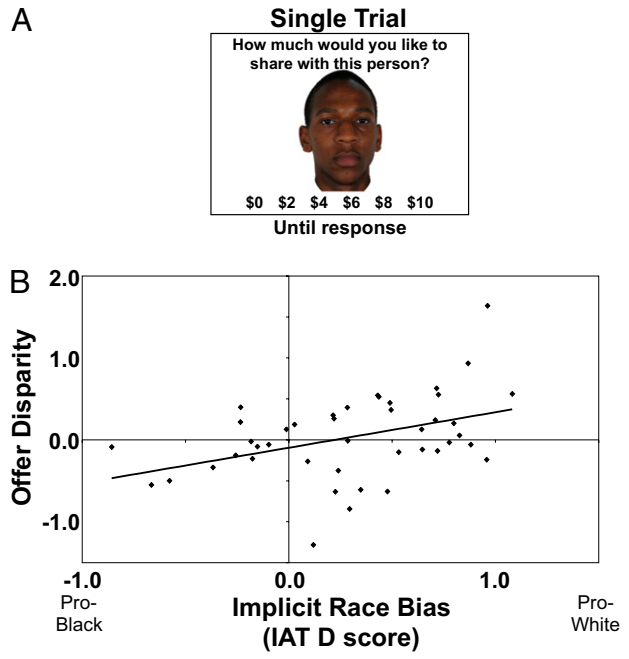

Fig. 2. Study 2: Individual differences in implicit race attitude correlate with race disparity in economic decisions. (*A*) Diagram of a single modified Trust Game trial. The stimuli were the same as in study 1. (*B*) Scatter plot showing each individual's score on the black/white, pleasant/unpleasant IAT and their black/white disparity in offers [Pearson's $r(41) = 0.4072$; $P = 0.0067$]. Offer Disparity = Mean(white offer *z* score) − Mean(black offer *z* score).

clusion methodology is discussed in *SI Text*; Fig. S3), indicating that participants had a slight overall aversion to risking their money [compare with loss aversion (31)]. There was no significant difference between mean offers to black (μ = $3.74, SD = $1.99) and white (μ = $3.75, SD = $1.72) partners for all participants [$t(42)$ $P = 0.688$, two-tailed paired Student's *t* test using the *z*-scored ratings data] or for white [$t(21)$ $P = 0.692$] or nonwhite [$t(20)$ $P = 0.311$] participants when considered separately. The mean IAT score was 0.29 (SD = 0.48; μ = 0.39, SD = 0.41 for white participants; μ = 0.20, SD = 0.53 for nonwhite participants; and no significant difference between the two). As in study 1, our subsequent analyses in study 2 focused on the relationship between individual differences in implicit race attitude and trust decisions involving monetary offers.

As with perceived trustworthiness, individual differences in implicit race attitude (IAT D score) predicted disparity in decisions to trust black and white partners. Individuals whose IAT scores reflected a stronger pro-white implicit bias were likely to offer more money to white partners than black partners, and vice versa. Using the same procedure as in study 1, we calculated each individual's black/white disparity in monetary offers and their subsequent ratings of trustworthiness. Individual differences in IAT score were significantly correlated with both offer disparity [all participants, $r(41) = 0.4072$, $P = 0.0067$; white participants only, $r(20) = 0.42$, $P = 0.055$; nonwhite participants only, $r(19) = 0.5$, $P = 0.021$; robust regression $\beta_{IAT} = 0.4$, $P = 0.012$; $\beta_{Participant\_Race} = -0.175$, $P = 0.23$; Fig. 2*B*] and subsequent rating disparity [all participants, $r(37) = 0.34$, $P = 0.036$; white participants, $r(19) = 0.145$, $P = 0.53$; nonwhite participants, $r(16) = 0.493$, $P = 0.038$; robust regression, $\beta_{IAT} = 0.42$, $P = 0.041$; $\beta_{Participant\_Race} = 0.052$, $P = 0.78$]. Again, the relationship between IAT scores and trust was robust [seen in 95% of random samples as small as 15 offers (5 offers from each race category); *SI Text*, Fig. S2] and independent of any relationship with standard measures of explicit race attitude, political identity, or participant race (stepwise regression analysis; Table 1).

Discussion

In two experiments with distinct populations and two different assessments of trust, our data demonstrate a surprisingly robust relationship between our evaluation of whether someone we do not know is trustworthy and our implicit bias with respect to their social group. We report and show greater trust in members of those groups toward whom we implicitly feel more favorable, and we do so independently of our explicit consciously accessible beliefs. In other words, our behavior is not driven solely by what we would consciously desire or intend.

Although we dedicate a portion of our results section to analyses of the role of participants' race in their ratings and offers, we do this only to demonstrate that their race does not account for our findings. We consider performance on the IAT to be a behavioral indication of general valenced associations with social groups, having little to do directly with one's own race. Although there can be relationships between one's own race and one's implicit biases, the susceptibility of implicit attitudes to experience (25, 26) suggests that such relationships emerge as a result of the shared experiences of those within a social group. Because continuous measures of implicit attitudes can access the effects of individuals' specific experiences, they offer much greater explanatory power than the categorical and subjective societal construct of race. Our findings that variations in performance on the IAT predict estimations of trustworthiness independent of participant race support a generalized mechanism whereby individuals' stored implicit associations with social groups can influence their conscious social decisions.

Our demonstration of the role of individual differences in shaping participants' trustworthiness estimations and trust decisions is a critical extension of previous work on trust (4, 5). Those studies report a high level of intersubject agreement concerning the trustworthiness of a given face, providing evidence that certain facial characteristics universally influence estimations of trustworthiness. Indeed, replicating those analyses, we found a high level of intersubject agreement for both ratings and offers in studies 1 and 2 (*SI Text*). That we find this agreement in addition to individual differences related to bias in implicit race attitude indicates that participants' estimations of trustworthiness were influenced by both the characteristics of the faces they viewed and their own implicit social biases. This suggests that future models of trustworthiness estimation would benefit from incorporating components that account not only for stimulus-driven effects contributing to intersubject agreement but also for observer effects (both implicit and explicit) that contribute to individual variation. In other words, to understand trust-based interactions fully, future research must examine not only the characteristics of the partner that make him or her trustworthy but the attitudes of the individual evaluating that partner's trustworthiness.

We state the result in terms of pro-white bias. This being a correlation, we could as easily have stated the result in the opposite direction (i.e., those with stronger pro-black bias were likely to find black faces/partners more trustworthy than white faces/partners). The IAT data in our experiments, as in previous studies (14), include a majority of individuals who showed pro-white bias (80%). For this reason, we report the data as we do, framed in terms of the pro-white bias of the majority of participants. We acknowledge that the relatively small number of participants with pro-black bias (i.e., IAT D < 0) in our sample may complicate inference about individuals at this end of the spectrum. We note that our study was designed to investigate the relationship between individuals' race-IAT scores and their rating and offer disparities rather than to make inferences at the group level. Nevertheless, the confirmatory results of the robust regression analyses, as well as the fact that most participants with

pro-black biases also had pro-black rating (5 of 6) and offer (9 of 12) disparities, give us confidence in the validity of our result across the range of possible IAT scores.

An interesting question concerns the specificity of the particular race IAT we used in predicting trust disparity. Are these findings limited to disparities in white vs. black trustworthiness estimations, or does the black/white IAT also predict race-related trust disparities for other social groups? We note that our experiment was designed specifically to investigate the relationship between implicit attitude and trust for black and white males. Other-race faces and partners were included to ensure that the dimension of interest was not inadvertently revealed to the participant. Even so, we examined the relationship between black/white implicit attitude and trust disparities for white vs. other-race and black vs. other-race faces and partners (*SI Text*). This analysis revealed that IAT scores were significantly predictive of white vs. other-race trust disparities and were somewhat less so for black vs. other-race trust disparities. This indicates that our findings may indeed generalize to other social groups and suggests that the particular IAT we used may be accessing a more generalized implicit social group bias than a specifically black/white bias.

It is important to note that our studies were purposely designed to obtain a relatively static snapshot of the participants' racial disparity in trustworthiness estimations (specifically the exclusion of outcomes in study 2). In more complicated situations, it is entirely likely that trustworthiness evaluations respond to context and previous outcomes in a dynamic and adaptive nature. Recent work demonstrating the malleability of implicit biases (through experience or situational manipulations; e.g., ref. 25) could be indicative of such flexible updating. A crucial next step toward understanding the relationship between implicit social biases and trustworthiness estimations would be to investigate the dynamic interaction between implicit bias and the outcomes of trust decisions.

The combination of information from neuroscience and psychology can serve as a powerful aid when constructing models of behavior and decision making. The behavioral link we have established between implicit race attitude and trust is consistent with a common neural substrate as suggested by previous work separately implicating the amygdala in both the expression of implicit race bias (19, 21) and trust evaluations (3, 4). This finding is correlative in nature, however, and, as such, represents only the initial step in demonstrating a causal relationship between implicit race attitude and trust. Future research should focus on demonstrating causality as well as obtaining direct evidence of overlapping and interrelated neural function to investigate the shared mechanism hypothesis. In addition, those interested in understanding the relationship between implicit attitudes and trust should make use of the large body of detailed knowledge concerning the neural mechanisms underlying fear learning and emotion processing in the amygdala to constrain putative models.

The current study is of specific interest to those examining trust and decision making, be it from a public policy, sociological, economic, or other point of view, because it identifies an independent factor that contributes to the decision process, namely, implicit social bias. In addition, these results are strong evidence that implicit measures predict real-world behaviors at the individual level and, as such, are a valuable tool for discerning the influence of processes that potentially lie outside our awareness. More generally, these data provide evidence that decisions we may believe to be consciously determined are, in fact, not entirely so and suggest that this may have a very real cost for individuals and society. In whom we trust is not only a reflection of who is trustworthy; it is also a reflection of who we are.

Methods: Study 1

Participants. Study 1 had 50 participants (27 women aged 18–39 y, mean age 22.2 ± 3.9 y; 31 white, 5 black, 7 Hispanic, 3 Asian, 4 multiracial). Study 2 had 59 nonoverlapping participants (40 women aged 18–39 y, mean age 21.7 ± 4.3 y; 31 white, 6 black, 4 Hispanic, 14 Asian, 4 multiracial), 2 of whom were excluded because of computer failure during data collection. In study 2, 52 of 59 participants returned later (mean = 14 d, range: 3–40 d) and provided trustworthiness ratings for their partners (as in study 1). All participants were recruited from the New York University community and surrounding area, provided informed consent, and were paid $10 per hour for participation. They were screened for English as their primary language and 10 or more years of residency in the United States. Race and ethnicity of participants were unrestricted to increase between-subject variance of IAT scores. Participants in study 2 received additional money based on the outcomes of their interactions in the modified Trust Game. All research reported here was approved by the University Committee on Activities Involving Human Subjects at New York University.

Stimuli. Three hundred eleven color pictures of forward-facing male faces with neutral affect (110 black, 110 white, and 91 other race[†]) were compiled from the Karolinska Directed Emotional Faces (32), the Eberhardt Laboratory Face Database, the Color Facial Recognition Technology Database from the National Institute of Standards and Technology, and the NimStim Face Stimulus Set (33). Faces were selected for picture quality, neutral expression, and hairstyles that were not clearly out of date. Eprime (Psychology Software Tools, Inc.), Psychtoolbox (34, 35), and MATLAB (MathWorks) were used for stimulus presentation in addition to paper surveys.

Procedure. Task order [trustworthiness rating task (study 1) or Trust Game (study 2), followed by the black/white, pleasant/unpleasant IAT and, finally, explicit measures] was fixed to minimize participants' awareness of the racial component of the study during completion of the main task. Study 2 participants were then asked to return later to complete the trustworthiness ratings task (*SI Text*). On completion of the experiment, all participants were fully debriefed as to the goal of the study and the nature of the IAT, and those in study 2 were debriefed concerning the deception involved in the Trust Game, in accordance with the guidelines provided by the University Committee on Activities Involving Human Subjects.

Trustworthiness Ratings. This task used a procedure similar to that of Todorov and colleagues (4). On each trial (Fig. 1*A*), participants saw a photograph of a face for 1 s and were then asked to rate how trustworthy that individual was on a scale from 1 (not at all trustworthy) to 9 (extremely trustworthy). Participants were assured their ratings would be anonymous and were asked to report their initial "gut impressions." The rating screen remained until participants responded using the number keys at the top of the keyboard. Trials were separated by 1 s. Participants saw three blocks of 97 faces each (total of 291: 100 black, 100 white, and 91 other race) with short self-paced breaks between blocks. Each face was shown only once, and the order of presentation was randomized.

Trust Game. Immediately after providing informed consent, participants were endowed with $30 in a room other than the experiment room. They were explicitly told that that money was theirs to keep and asked to put it with the rest of their money, wherever they kept it (e.g., wallet). They were then taken into the experiment room and were given a thorough briefing on how to play the game and with whom they would be playing. Specifically, they were told that they would be participating in real interactions with partners whom we had previously interviewed. They were then told that any money they shared would be quadrupled and then belonged to the partner, who had already made the decision either to share it with them (50/50) or to keep it all (full description of the instruction procedure is provided in *SI Text*; *SI Appendix A*). We emphasized that participants could either make sizeable amounts of money or lose the entire endowment in these interactions. Participants were told that their partners' faces were provided to help them get an idea of with whom they were playing and that three randomly selected trials would be realized at the end of the experiment (encouraging a focus on each individual interaction). After the instructions, participants were given a short written quiz to ensure task comprehension and any errors were discussed until the quiz could be completed correctly. Participants were fully informed concerning the design of the experiment, with the exception that they were led to believe the interactions were real.

On each trial (Fig. 2*A*), a face would appear in the center of the screen with the question "How much would you like to share with this person?" above the picture and the values "$0," "$2," "$4," "$6," "$8," and "$10"

[†]The "Other-race" group included individuals of Asian, Latino, and undetermined descent.

with their corresponding keys underneath. This display remained until participants responded. Participants did not see the outcome of each trial, and the next trial started 1 s later. As in the rating procedure, participants made 291 decisions (three blocks of 97 decisions) in random order. Finally, once the entire experiment was concluded, three randomly selected trial outcomes (randomly assigned "Share" or "Keep"; $P = 0.5$) were revealed and realized (i.e., participants either lost some of their endowment or received more money).

IAT. We administered a black/white, pleasant/unpleasant IAT using the procedure described by Lane et al. (36). Stimuli consisted of black and white faces (10 each, not used in the ratings/Trust Game portion of the experiment) and pleasant and unpleasant words (e.g., great, fantastic, terrible, awful). The order of congruent and incongruent blocks was randomly assigned, as was the hand assigned to the black or white category. Participants were reminded between blocks to go as fast as they could and that making some mistakes was acceptable. Participants' implicit race bias (IAT D score) was calculated using the algorithm described by Lane et al. (36).

Questionnaires. Participants completed questionnaires assessing explicit race biases, including the Modern Racism Scale (27), the Symbolic Racism Scale (28), and the Internal/External Motivation to Avoid Prejudice Surveys (29), as well as a set of explicit association indices (*SI Text*; Table S2). All these questionnaires were administered via computer in random order (question order within each scale was also randomized). In addition, we collected demographic measures, contact measures, participants' liberal/conservative affiliation (30), and a final questionnaire assessing participants' knowledge and beliefs about the study. These last surveys were administered on paper.

ACKNOWLEDGMENTS. We thank M. Perino and B. Capestany for assistance with data collection and analysis. This work was supported by grants from the MacArthur and Third Millennium Foundations.

1. Coleman JS (1990) *Foundations of Social Theory* (Harvard Univ Press, Cambridge, MA).
2. Todorov A (2008) Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Ann N Y Acad Sci* 1124:208–224.
3. Winston JS, Strange BA, O'Doherty J, Dolan RJ (2002) Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nat Neurosci* 5:277–283.
4. Engell AD, Haxby JV, Todorov A (2007) Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *J Cogn Neurosci* 19:1508–1519.
5. van't Wout M, Sanfey AG (2008) Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition* 108:796–803.
6. Pogrebin R (2008) Madoff Scandal, Jews Feel an Acute Betrayal. *NY Times,* US section, p A13. Available at http://www.nytimes.com/2008/12/24/us/24jews.html?em.
7. Delgado MR, Frank RH, Phelps EA (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat Neurosci* 8:1611–1618.
8. Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social-history. *Games Econ Behav* 10:122–142.
9. Todorov A, Engell AD (2008) The role of the amygdala in implicit evaluation of emotionally neutral faces. *Soc Cogn Affect Neurosci* 3:303–312.
10. Johnson ND, Mislin A (2008) Cultures of kindness: A meta-analysis of trust game experiments. Available at Social Science Research Network: http://ssrn.com/abstract=1315325.
11. Greenwald AG, Banaji MR (1995) Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychol Rev* 102:4–27.
12. Green AR, et al. (2007) Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *J Gen Intern Med* 22:1231–1238.
13. Caruso EM, Rahnev DA, Banaji MR (2009) Using conjoint analysis to detect discrimination: Revealing covert preferences from overt choices. *Soc Cogn* 27:128–137.
14. Nosek BA, et al. (2007) Pervasiveness and correlates of implicit attitudes and stereotypes. *Eur Rev Soc Psychol* 18:36–88.
15. Fazio RH, Jackson JR, Dunton BC, Williams CJ (1995) Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *J Pers Soc Psychol* 69:1013–1027.
16. Greenwald AG, Poehlman TA, Uhlmann EL, Banaji MR (2009) Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *J Pers Soc Psychol* 97:17–41.
17. Jost JT, et al. (2009) The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior* 29:39–69.
18. Phelps EA, LeDoux JE (2005) Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron* 48:175–187.
19. Phelps EA, et al. (2000) Performance on indirect measures of race evaluation predicts amygdala activation. *J Cogn Neurosci* 12:729–738.
20. Phelps EA, Banaji MR (2005) *Social Neuroscience: People Thinking About Thinking People* (MIT Press, Cambridge, MA), pp 229–243.
21. Stanley DA, Phelps EA, Banaji MR (2008) The neural basis of implicit attitudes. *Curr Dir Psychol Sci* 17:164–170.
22. Phelps EA, Cannistraci CJ, Cunningham WA (2003) Intact performance on an indirect measure of race bias following amygdala damage. *Neuropsychologia* 41:203–208.
23. Greenwald AG, McGhee DE, Schwartz JLK (1998) Measuring individual differences in implicit cognition: The implicit association test. *J Pers Soc Psychol* 74:1464–1480.
24. Banaji MR (2001) Implicit attitudes can be measured. *The Nature of Remembering: Essays in Honor of Robert G. Crowder*, eds Roediger HL, III, Nairne JS, Neath I, Surprenant A (American Psychological Association, Washington, DC), pp 117–150.
25. Blair IV (2002) The malleability of automatic stereotypes and prejudice. *Pers Soc Psychol Rev* 6:242–261.
26. Olsson A, Ebert JP, Banaji MR, Phelps EA (2005) The role of social groups in the persistence of learned fear. *Science* 309:785–787.
27. McConahay JB (1986) *Prejudice, Discrimination, and Racism* (Academic, San Diego).
28. Henry PJ, Sears DO (2002) The symbolic racism 2000 scale. *Polit Psychol* 23:253–283.
29. Plant EA, Devine PG (1998) Internal and external motivation to respond without prejudice. *J Pers Soc Psychol* 75:811–832.
30. Jost JT, et al. (2007) Are needs to manage uncertainty and threat associated with political conservatism or ideological extremity? *Pers Soc Psychol Bull* 33:989–1007.
31. Kahneman D, Tversky A (1979) Prospect theory—Analysis of decision under risk. *Econometrica* 47:263–291.
32. Lundqvist D, Flykt A, Ohman A (1998) The Karolinska Directed Emotional Faces—KDEF [CD-ROM] (Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet, Stockholm, Sweden).
33. Tottenham N, Borscheid A, Ellertsen K, Marcus DJ, Nelson CA (2002) Categorization of facial expressions in children and adults: Establishing a larger stimulus set. *J Cogn Neurosci* 14(Suppl):S74.
34. Brainard DH (1997) The psychophysics toolbox. *Spat Vis* 10:433–436.
35. Pelli DG (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spat Vis* 10:437–442.
36. Lane KA, Banaji MR, Nosek BA, Greenwald AG (2007) *Implicit Measures of Attitudes* (Guilford Press, New York), pp 59–102.