Casting a Wary Eye: Individuals Higher in Dispositional Distrust Demonstrate More Accurate Discrimination of Trustworthy and Untrustworthy Faces

Julianna R. Calabrese¹ · Mitch Brown² · Donald F. Sacco²

© Springer International Publishing 2016

Abstract Past research demonstrates that individuals are relatively accurate at discriminating trustworthiness from untrustworthiness utilizing facial cues alone, and that this capacity is augmented for those with activated self-protection threat concerns. In the current study, we predicted that individuals who are dispositionally more wary of trusting others (those scoring high in dispositional distrust) would be more accurate at discriminating trustworthy from untrustworthy faces. Participants viewed a series of trustworthy and untrustworthy faces and indicated whether each target was trustworthy or untrustworthy; participants then completed a general distrust inventory. Consistent with predictions, those higher in dispositional distrust demonstrated more accurate discrimination of trustworthy and untrustworthy faces. Additionally, higher dispositional distrust was associated with a higher criterion for reporting targets as trustworthy. Interestingly, the higher discriminability and criterion of more distrustful individuals seemed to be driven by a tendency to make fewer false alarms (i.e., decisions to categorize an untrustworthy face as trustworthy), but not at the expense of fewer hits (i.e., decisions to categorize a trustworthy face as such). Despite the necessity of trust for social affiliation, these results suggest greater dispositional distrust may facilitate the identification of favorable conspecifics for social exchange and poor social exchange partners to be avoided.

Keywords Distrust · Personality · Social perception · Trustworthiness

Cooperation is paramount to human survival. For social species to reap the benefits of cooperation, individual conspecifics must abide by principles of reciprocal altruism (Trivers 1971). Effective social living necessitates identification of trustworthy others to facilitate social cooperation as well as untrustworthy, potentially exploitative, others with the hopes of subsequently associating with the former and avoiding the latter (Rempel, Holmes, and Zanna 1985). Trustworthiness is one of the most desirable characteristics for interaction partners (Cottrell, Neuberg, and Li 2007), with people consistently placing greater value on traits indicating benevolence in various relationships (i.e., honesty, loyalty; Schwartz and Bardi 2001). Perceiving individuals as trustworthy consequently elicits greater cooperation, a prerequisite to forming social groups (e.g., Krumhuber et al. 2007). A suite of strategies to facilitate identifying veridical cues of trustworthiness in conspecifics would be adaptive to mitigate one’s vulnerability to exploitation.

Detecting Trustworthy Versus Untrustworthy Conspecifics

Humans can use several cues to determine conspecifics’ relative trustworthiness. For example, one can reflect upon observations of another’s past behavior to determine that person’s propensity to help or exploit as an index of potential future trustworthiness; if one recalls several occasions in which another has violated social contracts, that person might be deemed as generally untrustworthy (e.g., Chang et al. 2010). Additionally, individuals can use information from a third
party as an approximation of another’s relative trustworthiness. For example, research demonstrates that individuals who observe another person’s antisocial behavior feel compelled to share information about the actor with a potentially vulnerable person, which ultimately inhibits subsequent cooperation with the subject of this information (De Backer, Larson, Fisher, McAndrew, and Rudnicki 2016; Feinberg, Willer, Stellar, and Keltner 2012); such prosocial gossip is especially likely in individuals dispositionally higher in prosocial orientations. Although using information about another’s past behavior (whether obtained directly from interactions with that person or from informants) is critical for inferring another’s trustworthiness, individuals must often make such judgments of trustworthiness without this information, such as when meeting or interacting with someone for the first time (Delton, Krasnow, Cosmides, and Tooby 2011). Under these latter conditions, it would be adaptive to have sensitivity to other cues indicative of another’s relative trustworthiness.

Importantly, research indicates that individuals can infer another’s relative trustworthiness from facial cues. For example, research demonstrates that people associate wider faces, wider jaws, smaller eyes, and greater brow prominence with untrustworthiness (e.g., Haselhuhn and Wong 2011; Stirrat and Perrett 2010), and that people can infer others’ trustworthiness from brief exposure to faces (Slepian et al. 2012). Critically, aspects of facial appearance predict actual trustworthiness in interpersonal contexts (Stirrat and Perrett 2010). Such perceptions are also used to infer targets’ previous behavior and predict subsequent willingness to cooperate (Rezlescu, Duchaine, Olivola, and Chater 2012). Accurately inferring trustworthiness from facial cues is a critical perceptual adaptation as it allows for the identification of safe or threatening interaction partners prior to exploitation and without having to rely on third-party information, which itself may be inaccurate.

**Individual Differences in Trust and Social Perception**

A growing body of research has documented various personality characteristics and motives that predict the ability to accurately discriminate trustworthy from untrustworthy others. For example, acute activation of self-protection concerns leads men and women to demonstrate enhanced trust discrimination accuracy from faces (Young, Slepian, and Sacco 2012). Additionally, women with greater dangerous world beliefs are better at discriminating Duchenne smiles, which signal affiliation and cooperative intentions, from non-Duchenne smiles, which may be deceptive and mask underlying exploitative intentions (Sacco et al. 2016b). In the context of error management and social exchange theories (e.g., Cosmides and Tooby 1989; Haselton and Nettle 2006), it would have been historically more costly to assume someone is trustworthy when they are not (false alarm), than to assume someone is untrustworthy when they are actually trustworthy (miss). Whereas the latter decision may lead to missed cooperative opportunities, the former opens one up to potential exploitation. Thus, humans may have evolved to be relatively wary in trusting conspecifics, particularly novel conspecifics.

Nonetheless, research indicates individual differences in people’s tendency to trust others. For example, using a public goods game, researchers could classify participants into trust types: strong free-riders (i.e., those demonstrating low trust of others), conditional cooperators (those who trust others who display reciprocity), and strong cooperators (those demonstrating high trust toward others; Kurzban and Houser 2001). Recognizing differences in individuals’ ability to trust should thus indicate variability in thresholds used in determining whether a prospective conspecific is trustworthy. Subsequent trust game research comparing monozygotic with dizygotic twins revealed a significant genetic component to individual differences in people’s tendency to trust others (Cesarini, Dawes, Fowler, Johannesson, Lichtenstein, and Wallace 2008). Beyond assessing trait (dis)trust based on behavior in economic games, Yamagishi (1986) developed a personality questionnaire that captures individual differences in wariness with respect to trusting others (i.e., relative distrust). Importantly, such individual differences in distrust predict cooperative behaviors, including contributing to a public good in prisoner’s dilemma games (De Cremer, Snyder, and Dewitte 2001; De Cremer and Tyler 2007).

**Current Research**

Given that past research has identified individual differences in people’s tendency to trust/distrust others (e.g., Yamagishi 1986) and that people can infer another’s relative trustworthiness from facial cues (e.g., Slepian et al. 2012), the current research explored whether individual differences in distrust are related to the ability to discriminate trustworthy from untrustworthy others from facial cues. Specifically, we hypothesized that those with greater dispositional distrust of others would demonstrate enhanced ability in discriminating trustworthy from untrustworthy faces. Being able to identify trustworthy versus untrustworthy others would assist individuals in successfully identifying potential cooperative versus exploitative conspecifics, which would be especially critical to those higher in distrust of others. Additionally, we sought to determine if participants higher in dispositional distrust also set a higher threshold for identifying someone as trustworthy. Given that those with higher dispositional distrust might generally be more wary of conspecifics as a way to avoid exploitation, it is plausible that they set a higher threshold for identifying another as potentially trustworthy. To test these hypotheses, participants completed a face categorization task in...
which they viewed faces of trustworthy and untrustworthy targets while indicating whether each target was trustworthy. Following this task, participants completed a scale measuring individual differences in distrust, specifically, the extent to which they think others are generally untrustworthy.

**Method**

**Participants**

One hundred seventy-nine workers (83 men, 96 women; $M_{age} = 35.44$ years, $SD = 12.17$; 73 % Caucasian) participated through Amazon’s Mechanical Turk service in exchange for $0.35 (US). Participants first clicked a link taking them to the consent form; those interested in participating checked a box which advanced them to the survey while those uninterested in participating were instructed to close their browser window. One male participant did not complete the trust survey, making our final sample $N = 178$.

**Materials**

**Trustworthy-Untrustworthy Categorization Task** This task consisted of 34 Caucasian male faces (17 trustworthy, 17 untrustworthy) displaying neutral facial expressions. These faces were normed by previous research, such that half of the faces were consensually judged as high in trustworthiness and the other half were consensually judged as low by an independent sample of participants (see Slepian et al. 2012). Participants viewed each face once, in randomized order; they were asked to indicate whether they believed each face was trustworthy or untrustworthy. Thus, accuracy in the context of this paradigm is the extent to which participants’ judgments of target trustworthiness and untrustworthiness are consistent with normative judgments of these same targets. Consistent with past research exploring trust discrimination accuracy, we adopted a signal detection framework for analyzing participants’ responses (Sacco, Merold, Lui, Lustgraaf, and Barry 2016; Young, et al. 2015). We coded HITs as trials in which participants categorized a trustworthy face as trustworthy and coded false alarms (FAs) as trials in which participants categorized an untrustworthy face as trustworthy. We used HITs and FAs to compute participants’ d-prime ($d'$), where higher values indicate greater ability to accurately discriminate trustworthy from untrustworthy faces, as well as each participant’s criterion ($β$), where higher values indicate participants’ greater requirement of evidence to report a face as trustworthy.  

1 As is common in signal detection analyses, adjustments to the data were made to address the problem of empty cells: 0 % was adjusted to 5 %, and 100 % was adjusted to 95 % (see Hugenberg et al. 2007). Alternate adjustments yielded nearly identical results.

**Trust Scale** To measure individual differences in trust, we utilized a scale developed by Yamagishi (1986). This five-item scale includes statements such as “Most people tell a lie when they can benefit by doing so,” and “Those devoted to unselfish causes are often exploited by others.” Participants reported their (dis)agreement with each statement, utilizing a 5-point Likert-type scale (1 = strongly disagree; 5 = strongly agree) with higher values indicating greater dispositional distrust. The scale had acceptable reliability ($α = .64$).

**Procedure**

Following informed consent, participants first completed the trust face discrimination task followed by the trust scale, and finally a demographic form (i.e., age, gender, race). Then, participants were redirected to a debriefing screen before receiving a code to redeem for study payment.

**Results**

**Preliminary and Correlational Analyses**

Prior to conducting correlational analyses, we ran two custom ANCOVAs with participant gender as a between-subject independent variable and individual differences in distrust as a covariate, one for $d'$ and one for $β$. These models allowed us to test for any effects of gender. Neither model produced main effects of gender nor did gender interact with individual differences in distrust (all $ps > .233$). As such, we excluded gender from all primary analyses reported below. Importantly, men ($M = 3.17$, $SD = .60$) and women ($M = 3.13$, $SD = .69$) did not differ in their levels of dispositional distrust, $t(176) = .418, p = .676, d = .06$.

Consistent with our hypothesis, there was a significant correlation between distrust and $d'$, $r(176) = .153, p = .042$, such that those higher in dispositional distrust were better able to discriminate trustworthy from untrustworthy faces. Additionally, there was a significant correlation between distrust and $β$, $r(176) = .237, p = .001$, such that those higher in dispositional distrust had a higher criterion for reporting faces as trustworthy.

**Decision Processes Related to Higher Criterion and Discriminability**

Because those higher in dispositional distrust demonstrated both a higher criterion for categorizing a given face as trustworthy, as well as demonstrated an increased ability to discriminate trustworthy from untrustworthy faces, we sought to determine what aspects of participants’ decision-making might be driving these effects. Specifically, individuals’ criterion and sensitivity are derived from their proportions of HITs.
(correctly identifying a trustworthy face as trustworthy) and FAs (incorrectly identifying an untrustworthy face as trustworthy). In the context of the current trust discrimination task, it would be most adaptive if increased sensitivity and criterion among those higher in distrust were driven by a reduced tendency toward FA decisions, without a reduction in HITs. Specifically, more distrustful participants, under these conditions, would be successful in avoiding poor social exchange partners (untrustworthy targets) without missing out on true cooperative opportunities (trustworthy targets). As such, we correlated participants’ dispositional distrust with their proportion of HITs (total number of trials in which a participant categorized an untrustworthy face as trustworthy divided by the total number of trustworthy face trials) and proportion of FAs (total number of trials in which a participant categorized an untrustworthy face as trustworthy divided by the total number of untrustworthy face trials). Higher dispositional distrust was not significantly correlated with HIT rate, \( r(176) = -.127, p = .092 \), suggesting that those higher in dispositional distrust do not demonstrate a compromised capacity to successfully identify trustworthy others (i.e., those higher in dispositional distrust do not differentially mis-categorize more trustworthy targets as untrustworthy). However, higher dispositional distrust was significantly related to fewer FAs, \( r(176) = -.293, p < .001 \). Thus, those higher in dispositional distrust are significantly less likely to accidentally categorize an untrustworthy target as trustworthy. Thus, the higher criterion and enhanced trust discrimination accuracy displayed by participants higher in distrust seem adaptive: it leads to fewer false alarms, thereby reducing potential exploitation by untrustworthy conspecifics, while still allowing for accurate identification of conspecifics who are, in fact, trustworthy, thereby facilitating true cooperative opportunities.

Discussion

These results indicate that participants who are dispositionally less trusting of others demonstrate greater accuracy when tasked with categorizing faces as trustworthy or untrustworthy. Such perceptual acuity would be adaptive as these individuals should be more judicious in their conspecific selection to prevent exploitation. Interestingly, participants higher in dispositional distrust also had a higher criterion, such that they required more evidence in order to categorize any face as trustworthy. Importantly, these increases in criterion and discriminability with respect to identifying trustworthy and untrustworthy targets were due to a reduced tendency to mis-categorize untrustworthy targets as trustworthy, but not at the expense of failing to identify trustworthy targets as trustworthy. Thus, those higher in distrust are able to identify trustworthy conspecifics (favorable social exchange partners) without simultaneously mistakenly categorizing untrustworthy targets (potentially exploitive conspecifics) as trustworthy.

Our findings are consistent with other research indicating how dispositionally higher wariness of others’ intentions enhances accuracy for facially communicated trustworthiness and affiliative intent. Indeed, women’s dangerous world beliefs (Sacco et al. 2016b) and activated self-protection motives (Young et al. 2015) enhance one’s ability to identify safe interaction partners based on veridical cues of trustworthiness (e.g., facial structure, facial affect), but these findings may not necessarily address general wariness of conspecifics beyond concerns over threats to physical safety.

Our results extend previous findings by demonstrating this acuity at the dispositional level of a more ubiquitous motivation. That is, whereas self-protection motives specifically elicit perceptions based on concern about immediate physical safety, the current study indicates that wariness of untrustworthy others may extend to other motivational states. For example, the trust questionnaire partially deals with the role of social welfare in wealth redistribution (Yamagishi 1986), thus making it seem plausible that participants’ indication of distrust in this measure may be tied to their concern of losing resources. This may suggest that a facet of distrust concerns one’s desire to protect their resources or being exploited financially. Those who are more wary about associating with someone who could exploit them financially would thus benefit from recognizing those whom they view as more likely to be a resource threat (i.e., untrustworthy targets).

Furthermore, there were no participant sex differences found in the current study; however, past research has found that individual differences in women’s (but not men’s) dangerous world beliefs predict an enhanced ability to discriminate cooperative (targets displaying Duchenne smiles) from potentially exploitive targets (targets displaying non-Duchenne smiles). Though one might expect female participants to report greater accuracy in identifying exploitative conspecifics due to the physical disadvantages imposed by sexual dimorphism, the context of this study might have resulted in roughly equivalent accuracy across sexes. In general social groups, both men and women are at risk of being exploited by untrustworthy others; thus, a sex difference would not have emerged if the same environmental pressures presented themselves to men and women equally across human’s evolutionary history. Indeed, men and women reported similar levels of dispositional distrust, which suggests that exploitation concerns may be similar for both men and women.

Though we found predictive power in the Trust Scale (Yamagishi 1986), one limitation of the current study pertains to some of the psychometric properties of this scale. First, this scale only contained five items. Short scales have been criticized for both their lower accuracy in tapping a construct and diminished predictive validity. However, they are also
effective in that they can reduce obstacles posed by longer measures, thus providing a quicker index of the desired construct (Maloney, Grawitch, and Barber 2011). Additionally, the wording of some items from the distrust scale may have confused participants, such as one item being double-barreled, which may explain why the scale only demonstrated moderate internal consistency ($\alpha = .64$); however, individual item analysis revealed that no single question was responsible for the moderate reliability documented in the current study. Future studies may include additional scales, such as the General Trust Scale (Yamagishi & Yamagishi, 1994), to determine the generalizability of these findings, and how dispositional trust toward others may work in opposition to distrust.

**Future Directions**

There are several outlets for future directions in light of our current findings. First, though using male faces has robust implications, given how it was historically critical to identify untrustworthy men for either sex, it may prove similarly fruitful to consider perceptions of trustworthiness as a function of target sex. Perhaps male faces communicate trustworthiness differently from female faces, and that perception may ultimately be moderated by participant sex. Individuals may perceive trustworthiness differently in male compared to female faces. Less trustworthy men may provide greater intrasexual competition to other men and greater threat of exploitation to women in terms of physical harm or extra-pair relations (e.g., Sacco et al. 2016a). Untrustworthy female faces may similarly provide a cue to women as potential intrasexual rivals to make it adaptive for women to identify exploitative women who may limit mating opportunities, which may clarify findings suggesting women are less likely to cooperate in prisoner’s dilemma games with female partners (Balliet, Mulder, and Van Lange 2011). Given these potential bases for differences in accurately identifying a trustworthy conspecific, future research should ultimately identify the crux of these trustworthiness perceptions and if there are different reasons for men and women to perceive trustworthiness, especially based on the sex of a target. Future studies should consider norming female faces for trustworthiness to determine how social perceptions are influenced by target sex.

Given that we can perceive trustworthiness from an individual’s face, such perceptions may ultimately be able to predict interpersonal behavior in response to social targets. Indeed, previous research has demonstrated that dispositional distrust inhibits cooperative behavior and that facially communicated trustworthiness augments willingness to engage a social target (e.g., De Cremer et al. 2001; Rezlescu et al. 2012). Our findings would likely synthesize with previous work to offer an integrated model of dispositional distrust’s influence on social perception that ultimately predicts cooperative behavior. This could be accomplished through a prisoner’s dilemma game or similar cooperative tasks.

It is worth noting that in the current study, the stimuli were trustworthy and untrustworthy not based on their behavior, but because prior research indicated that the faces possessed characteristics that elicited consensual judgments of perceived trustworthiness or lack thereof. Future research should determine if dispositional distrust enhances the ability to identify others who actually have a history of being trustworthy or untrustworthy. For example, past research finds that people demonstrate enhanced memory for the faces of cheaters relative to cooperators (Mealey, Daood, and Krage 1996). Our results suggest that beyond sensitivity to facial cues that people associate with perceived trustworthiness or untrustworthiness, those higher in dispositional distrust should perhaps demonstrate enhanced cheater detection ability, such as enhanced memory for the faces of actual cheaters.

Other individual differences in interpersonal wariness may provide similar results while providing a behavioral analog for perceived trust. For example, reciprocation wariness has previously predicted greater negative attitudes toward uncooperative others that may serve as a robust predictor for both perceptual and behavioral outcomes. A component of humans’ prosociality is their expectation of reciprocal cooperation from conspecifics; those who expect reciprocity from their conspecifics would thus be more willing to behave prosocially without an initial prompt (e.g., Cialdini, Kallgren, and Reno 1991). However, individual differences exist in terms of willingness to initiate prosociality; some individuals are more wary of others’ intentions and have a greater fear of being exploited by conspecifics. Previous research indicates that individuals high in dispositional reciprocation wariness inhibit social relationships and are less open to engaging with others. Importantly, dispositionally warier individuals have expressly more negative attitudes toward uncooperative behavior (Cotterell, Eisenberger, and Speicher 1992). Future studies concerning reciprocation wariness should consider how this wariness heightens perceptual acuity toward facially communicated trustworthiness while also eliciting divergent behavioral patterns following perceptions of trustworthiness (or lack thereof). Specifically, the accurate identification of a trustworthy face as trustworthy should subsequently predict greater willingness to cooperate with that target in interdependent tasks and greater favorability. Conversely, identifying an untrustworthy face as untrustworthy should result in greater reluctance to cooperate and greater derogation.

**Conclusion**

This study provides novel evidence for how individual differences in distrust toward others influence social perception. Specifically, we found that more distrustful people
demonstrated greater accuracy in discriminating between trustworthy and untrustworthy faces and set a higher criterion to categorize a given face as trustworthy. Most importantly, this increased discriminability and criterion with respect to categorizing trustworthy and untrustworthy faces seemed to be driven by a tendency to make fewer detrimental categorizations of untrustworthy targets as trustworthy, without displaying a detrimental tendency to accidentally categorize actual trustworthy targets as untrustworthy. Although distrust may inhibit individuals from forming bonds with others, this inhibition appears only to prevent indiscriminate bonding that would open these individuals up to potential exploitation. Thus, distrust is an adaptive trait to ensure optimum group functioning.

References


