Scuola IMT Alti Studi Lucca

When faces shape places: person trustworthiness influences contextual threat

Questa è la versione preprint della seguente opera:

Original
When faces shape places: person trustworthiness influences contextual threat / Mattavelli, S., Masi, M., Brambilla, M., Fiamberti, G.C.. - In: COGNITION & EMOTION. - ISSN 0269-9931. - (2026). [10.1080/02699931.2026.2743254]

Availability:
This version is available at: 20.500.11771/45259

Publisher:
Taylor and Francis

Published
DOI:10.1080/02699931.2026.2743254

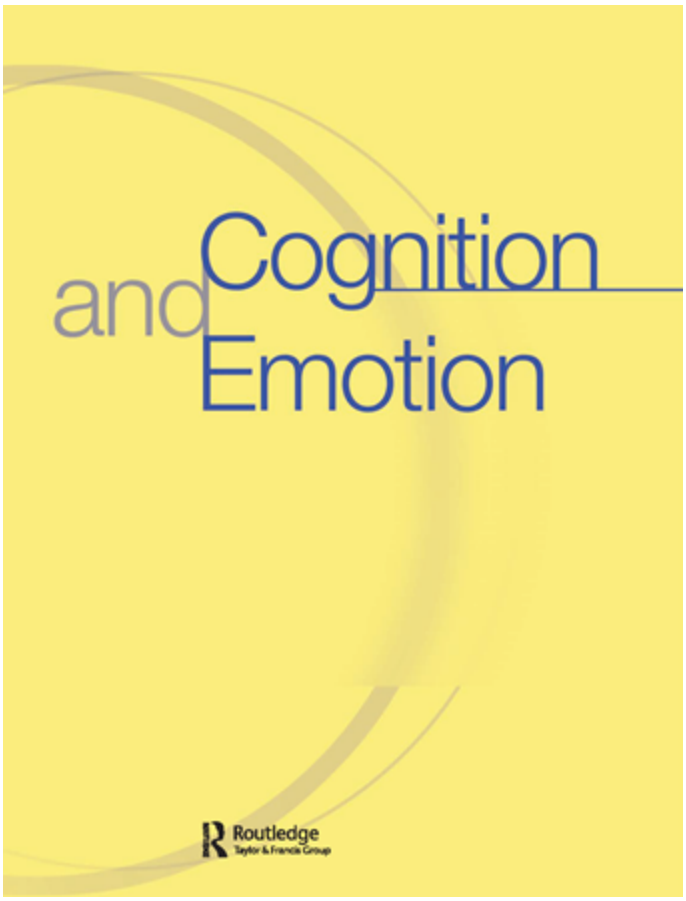
Terms of use:

This publication is made accessible in accordance with the terms for deposit in the institutional repository, as defined by the IMT School for Advanced Studies Lucca's Open Access Policy. (https://library.imtlucca.it/sites/default/files/regolamento-policy-open-access-imtlib_0.pdf).

Si prega di consultare le pagine informative dell'editore relative alle politiche di autoarchiviazione.

(Article begins on next page)

10 November 2026



When Faces Shape Places:Person Trustworthiness Influences Contextual Threat

Journal:	*Cognition and Emotion*
Manuscript ID	CEM-FA-86.26.R2
Manuscript Type:	Full Article
Date Submitted by the Author:	25-Sep-2026
Complete List of Authors:	Mattavelli, Simone; University of Milan–Bicocca, Psychology Masi, Matteo; University of Milan-Bicocca Brambilla, Marco; University of Milan-Bicocca, Department of Psychology Fiamberti, Gaia Carlotta; Scuola IMT Alti Studi Lucca
Keywords:	face-context integration, conditioning, relational reasoning, inferential process, trustworthiness


SCHOLARONE
MANUSCRIPTS

When Faces Shape Places:

Person Trustworthiness Influences Contextual Threat

Simone Mattavelli, Matteo Masi, Marco Brambilla

Università Milano-Bicocca

Gaia Carlotta Fiamberti

Università Milano-Bicocca, Scuola IMT Alti Studi di Lucca

For Peer Review Only

RUNNING HEAD: Trustworthiness and Contextual Threat

Corresponding author: Simone Mattavelli, University of Milano-Bicocca, Department of Psychology, 1, Piazza dell'Ateneo Nuovo, 20126, Milan, Italy.

E-mail: simone.mattavelli@unimib.it

Abstract

Social perception rarely occurs in isolation: people and places mutually shape one another's meaning. Prior research has shown that threatening contexts can bias impressions of facial trustworthiness, suggesting that evaluative meaning flows from the environment to the person. Here, we consider the reverse causal direction, testing whether individuals' trustworthiness can alter evaluations of environmental threat. Across three preregistered experiments (total N = 367), participants judged the threat level of ambiguous visual contexts that included individuals varying on perceived trustworthiness. Experiment 1 showed that untrustworthy faces increased perceived contextual threat relative to neutral or absent faces, an effect that correlated with attributing danger to the target individual. Experiment 2 revealed that this effect was contingent on perceiver perspective: untrustworthy faces heightened threat only when participants evaluated the context from their own viewpoint, not when adopting the target's perspective. Experiment 3 showed that this face-to-context influence emerged when faces varied in trustworthiness but not competence, indicating that it does not reflect a general negativity bias. Together, these findings establish face–context integration as a process in which social meaning propagates dynamically between person and environment, revealing a fundamental mechanism of relational meaning construction in social cognition.

Word count of abstract: 194

Keywords: face-context integration; conditioning; relational reasoning; inferential process; trustworthiness.

For Peer Review Only

When Faces Shape Places:

Person Trustworthiness Influences Contextual Threat

Imagine walking alone through a poorly lit parking garage late at night. The setting feels ambiguous, neither clearly safe nor overtly dangerous. Suddenly, you notice a person standing nearby. Something about their face feels cold, untrustworthy. Almost instantly, the space around you seems more menacing. But what exactly has changed? This scenario reflects a key process in social cognition: integrating facial and contextual information. Previous research has shown that threatening contexts can bias judgments of trustworthiness from faces (for a review, Brambilla et al., 2024). Yet this focus assumes that social meaning flows only from the environment to the person. Extending prior research, here, we explore whether the reverse also holds: Can the perceived trustworthiness of a person influence how threatening a context feels? Drawing on theories of relational reasoning, we propose that face–context integration reflects a broader principle of social meaning construction, in which perceivers interpret people and contexts not as independent cues but as parts of an integrated relational structure. From this perspective, evaluative meaning should not be constrained to flow from context to face but should also be able to flow from face to context. Across three experiments, we tested (i) whether untrustworthy individuals heighten perceived contextual threat, (ii) whether this effect depends on the perceiver's perspective, and (iii) whether it is specific to traits related to physical threat (i.e., trustworthiness) or extends to other social traits.

Contextual threat and facial trustworthiness

Social perception is inherently contextual. Rarely, we encounter individuals in isolation; instead, our impressions unfold within broader environments that subtly or overtly shape how we evaluate others. Hence, a person's facial expression may be interpreted differently depending on whether they appear in a sunny park, a corporate boardroom, or a dimly lit alley. This process, known as face–context integration, describes how perceivers incorporate contextual cues into judgments of social meaning from faces (Aviezer et al., 2011; Barrett & Kensinger, 2010; Masuda et al., 2008; Wieser & Brosch, 2012). Face-context integration reflects a broader principle in social cognition: the

For Peer Review Only

meaning of a stimulus is not fixed but emerges through its situational embedding. Research on emotion recognition nicely illustrates this principle: observers identify facial expressions more accurately when faces are presented within emotionally congruent contexts. For instance, disgusted faces are recognized more easily when embedded in scenes containing garbage or bodily injury rather than neutral backgrounds (Aguado et al., 2019; Righart & De Gelder, 2008). Such findings highlight that perceivers dynamically combine facial and contextual information to resolve ambiguity in complex social environments.

Among face-based judgments, trustworthiness stands out as particularly sensitive to contextual threat. Trustworthiness is one of the most perceptually salient traits that people readily infer from faces (Todorov et al., 2008; Todorov et al., 2009; Willis & Todorov, 2006). Although such impressions are subjective and often inaccurate (Foo et al., 2022; Rule et al., 2013) they exert powerful social consequences. For example, untrustworthy-looking individuals receive harsher sentences (Wilson & Rule, 2015), while trustworthy-looking hosts are more likely to rent out their apartments (Ert et al., 2016). Importantly, trustworthiness judgments are not determined solely by facial features but are contextually malleable. A growing body of research suggests that threatening environments systematically disrupt perceived facial trustworthiness (Brambilla et al., 2018; Brambilla et al., 2021). For instance, when neutral or ambiguous faces are presented in visually threatening environments (e.g., dark alleys, scenes of violence) or accompanied by threatening sounds (e.g., sirens, screams), they are consistently rated as less trustworthy than when the same faces are shown in neutral or merely unpleasant (but non-threatening) contexts (Brambilla et al., 2018; Brambilla et al., 2021). Moreover, this effect appears functionally specific: face-context integration appears stronger under the untrustworthiness-threat pairing than under other trait-emotion combinations (i.e., extraversion and happiness, Mattavelli et al., 2022). This selectivity suggests that threat and untrustworthiness form a tightly coupled evaluative system, likely rooted in adaptive goals of identifying danger and avoiding exploitation (Brambilla & Leach, 2014; Cosmides & Tooby, 1992; Fiske, 1992). Yet despite this rich literature, existing models of face-context integration implicitly

For Peer Review Only

assume that meaning flows in one direction – from the environment to the face. If threat cues in a context can make a neutral face appear untrustworthy, might the reverse also occur; in other words, could an untrustworthy face make a neutral context appear threatening? Addressing this question requires a theoretical framework capable of explaining why evaluative meaning might flow in both directions. Recent work suggests that context-to-face integration relies on relational reasoning, whereby observers infer the causal and social relations linking individuals and their environment (Mattavelli et al., 2023a, 2023b; Brambilla et al., 2024). From this perspective, evaluative meaning should also flow from face to context whenever faces provides informative basis for interpreting an otherwise ambiguous environment.

A relational view of face-context integration

The effects of context on trustworthiness judgments cannot be fully explained by simple associative mechanisms. Recent research suggests that face–context integration involves a more cognitively elaborated process of relational inference, whereby perceivers interpret the likely causal and social roles of individuals within a scene (Brambilla et al., 2024; Mattavelli et al., 2023a, 2023b). Observers do not merely react to the presence of a threatening environment by downgrading their trust of anyone perceived within it. Instead, they actively evaluate the role of the person in the surrounding background, drawing inferences about whether that person is likely to be the source of the threat, a victim, or a bystander. For example, trustworthiness judgments plummet when a face appears in a violent setting that implies the person may be a perpetrator; however, when cues like emotional expressions (e.g., fearful expression) or verbal labels suggest the same individual is a victim, trustworthiness may increase, even though the context does not change (Mattavelli et al., 2023a, 2023b). This relational interpretation aligns with a broader shift in the conditioning literature. Traditional evaluative conditioning accounts treat affective transfer because of repeatedly pairing two stimuli which should establish a unidirectional mental association between a conditioned stimulus (CS) and an unconditioned stimulus (US) (De Houwer, 2007). However, a propositional approach argues that conditioning also depends on the reasoning about the relationship between stimuli such

For Peer Review Only

that whether the CS causes, prevents, or merely co-occurs with the US (De Houwer, 2014, 2018, De Houwer et al., 2020). Indeed, a good deal of work has shown that relational information can reverse or amplify conditioning effects (e.g., Fiedler & Unkelbach, 2011; Förderer & Unkelbach, 2012; Moran & Bar-Anan, 2013). For example, Förderer and Unkelbach (2012) found reversal in conditioning effects when the CS disliked (vs. liked) the US stimulus.

Together, these findings suggest that affective meaning is not simply transferred from one element to another but constructed through the perceived relations between them. This insight bridges conditioning and social perception: in both domains, relational cues shape how social meaning flows across linked representations. Building on this observation, the present research extends this logic to a person-to-environment transfer.

Relational meaning construction

Face–context integration provides a unique window into how social meaning can flow bidirectionally between stimuli. In classic evaluative conditioning paradigms, one stimulus typically carries the valence (e.g., the US) and the other is neutral (e.g., the CS). In face-context integration, instead, the roles of face and context as CS or US are not fixed. Both elements can be affectively meaningful or ambiguous, and their assignment as source or target of evaluative meaning shifts dynamically. Depending on how the scene is interpreted, the face can be perceived as the source of danger, with the context becoming the target of judgment, or the other way around. This fluid attribution of roles marks a key feature: in relational face-context integration, evaluative meaning does not simply "flow" from one stimulus to another. Instead, it is jointly constructed, with the social function of each element shifting according to the perceiver's relational inferences. A relational perspective therefore conceives face–context integration as a bidirectional reasoning process, in which the interpretation of each element is contingent on its perceived relationship to the other. To exemplify, consider the parking garage late at night described earlier with a person standing nearby. If the person carries facial features that suggest malevolence, they might heighten the perceived

For Peer Review Only

danger of the context. Conversely, if the individual's face appears harmless and trustworthy, it might turn the potentially threatening situation into an innocuous scenario.

Prior work has revealed bidirectional influences between emotional expressions and situational interpretation (Hess et al., 2017, 2020), showing that observers use facial cues to infer the meaning of events and situational cues to refine emotion judgments. However, this research has almost exclusively focused on transient, goal-oriented scenes (e.g., moral transgressions, competition) rather than on trait-based evaluations such as trustworthiness. Here, we test whether the perceived trustworthiness of a target person can actively shape the threat value of their surrounding visual context. Specifically, we examine (a) whether untrustworthy faces heighten perceived contextual threat, and (b) whether this effect reflects perceivers' tendency to construe the face as the source of potential danger. Demonstrating this "face-to-context" influence would provide additional support to relational reasoning as a general mechanism of social meaning construction.

The present Research

Three experiments investigated whether and how perceived facial trustworthiness can influence evaluations of environmental threat. In Experiment 1, participants rated potentially threatening scenes that either featured an untrustworthy face, a neutral face, or no face at all. We hypothesized that untrustworthy faces would induce higher contextual threat, as opposed to when either neutral faces or no faces were presented. Experiment 2 examined whether the hypothesized effect depends on the perceiver's perspective, predicting that untrustworthy faces would heighten contextual threat only when participants judge the context from their own perspective, not the target's. Across both experiments, we also expected that higher contextual threat would be associated with a greater tendency to attribute a dangerous role to untrustworthy faces. Experiment 3 assessed whether the effect arises uniquely from trustworthiness or reflects a broader negativity bias. Using a conditioning paradigm, participants first learned to associate faces with either moral (trustworthy) or immoral (untrustworthy) behaviors, or with competent or incompetent behaviors. We expected that only faces paired with immoral behaviors would increase perceived contextual threat, whereas

For Peer Review Only

incompetent faces, despite being negative, would not. Moreover, we expected trustworthy faces would attenuate perceived contextual threat compared to competent faces. In line with prior research on face-context integration (but see Mattavelli et al., 2023a, Experiment 1c), all studies focused on male faces. Thus, the extent to which the present findings generalize to female faces remains an empirical issue (see the General Discussion).

Open science

All the studies were preregistered before data collection (Experiment 1: https://osf.io/4zxfr/overview?view_only=db31b5d1663f42a5997ce3c53e29246a; Experiment 2: https://osf.io/c7naq/overview?view_only=3bcdbe7574b84419b4d626da1dcbac48; Experiment 3: https://osf.io/8nqmp/overview?view_only=6f9e85a73357425a9af2be2fc874afe0).[1] We reported all data exclusions, planned analyses, and deviations from the pre-registrations. Materials, data, and analysis scripts are available at OSF (Experiment 1: https://osf.io/vjde3/overview?view_only=0dbaf5af41a0443fbc25f8c32abc11a3; Experiment 2: https://osf.io/y7689/overview?view_only=37dc928027f2413fad620074e77b108c; Experiment 3: https://osf.io/pbc8x/overview?view_only=d83c852526f04f63bbd741da65987cef).

All the experiments received formal approval from the departmental Ethics Committee.

Experiment 1

In the first experiment, we provide preliminary evidence that the presence of an untrustworthy face (vs. neutral face vs. no-face) leads observers to perceive the context in which the face is embedded in as more threatening. We also measured attributions of threatening intentions to the untrustworthy vs. neutral faces in the contexts. Perceived contextual threat due to untrustworthy vs. neutral faces should be positively related to the perception of untrustworthy vs. neutral faces as dangerous agents.

Method

[1] The OSF preregistration pages appear empty and contain only general information (e.g., study description, overview, and time stamp). However, the preregistered protocols for each study are available as detailed PDF documents within each study's OSF project, and their creation dates correspond to the preregistration time stamps.

For Peer Review Only

We employed a full-within design (face type: untrustworthy vs. neutral vs. no-face). Participants rated the threat of the context using a 7-point scale.

Sample Size Determination and Participants

The analysis was designed to ensure sufficient power to detect a significant effect of face type on judgments of contextual threat. More specifically, we aimed to achieve adequate power for all relevant contrasts. To determine the required sample size, we conducted a Monte Carlo simulation using SuperPower in R (Lakens & Caldwell, 2021). We employed a 3-level within-subject ANOVA with face type as the independent variable (untrustworthy face vs. neutral face vs. no face). We assumed two small effects (*Cohen's d* = 0.25) for the contrasts between no face and neutral faces (effect 1) and between neutral and untrustworthy faces (effect 2), as well as a medium effect (*Cohen's d* = 0.50) for the contrast between no face and untrustworthy faces (effect 3). With α set at .05, our simulation (2,000 iterations) indicated that a sample of 130 participants would provide: (i) 81% power to detect effect 1; (ii) 81% power to detect effect 2; (iii) 100% power to detect effect 3. Accordingly, we planned to slightly oversample and recruited at least 140 participants via Prolific Academic. In all Experiments, we included only those who speak Italian, with an approval rate of at least 90% and 100 previous submissions on Prolific and in Experiments 2 and 3, we excluded those who took part in the previous experiment.

We pre-registered a sequential analysis approach (Lakens, 2014) with a single interim analysis. Using the Pocock boundary to adjust the alpha level, we conducted an interim analysis after 70 participants had completed the study. Since one of the critical tests did not meet the adjusted significance threshold of $p < 0.0294$, we proceeded with data collection, ultimately obtaining complete responses from 147 participants (58 women, 87 men, 2 other genders; $M_{age} = 31.24$, $S_{Dage} = 9.61$).

Stimuli

Faces. We employed 16 male face identities displaying neutral expressions borrowed from the Chicago Face Database (Ma et al., 2015), of which according to the normative data 8 faces were

For Peer Review Only

perceived as untrustworthy (M = 2.59 [2.31; 2.82]; SD = 0.18) and 8 were perceived as neutral on the same dimension (M = 3.75 [3.59; 3.92]; SD = 0.10); (see Appendix C for the exact stimuli).

Contexts. In a pilot test (N = 39, 21 women and 1 non-answered, M_{age} = 31.13, SD_{age} = 10.68) we pretested context images taken from Mattavelli et al. (2022). Participants evaluated each image on a 7-point Likert scale, indicating how threatening the image was (1 = not at all threatening; 7 = extremely threatening). The images did not imply the presence of human actions and depicted environments such as car parks, streets, and city parks. Overall, the evaluations ranged from 1.46 to 5.51. We selected 16 stimuli (M = 4.21, SD = 0.76), in a range between 3.26 and 5.51. This was done to make sure that context stimuli could elicit threat when paired with a face, without being too extreme on this dimension (see OSF link for the full set of stimuli).

Procedure

The experiment was created in Millisecond's Inquisit. Participants gave consent to participate and provided demographics, then received the instructions. They viewed a series of scenarios displayed on a screen. For each scenario, they were asked to imagine themselves in the depicted setting. Some scenarios were presented alone, while others included a neutral or untrustworthy face at the center of the scene (see Figure 1 for an example). Participants were instructed to rate how threatening or dangerous each scenario appeared, using a scale ranging from 1 (not at all threatening) to 7 (extremely threatening). Each of the 16 contexts appeared three times[2] with each of the 8 neutral faces, 8 untrustworthy faces, and alone, leading to a total of 48 trials. Face identities were randomly paired with contextual scenes, preventing systematic associations between specific faces and contexts. Importantly, participants were asked to base their evaluation solely on the scenario itself and to disregard the presence of any face when making their judgments.

After the first task, participants were exposed again to face-context random pairs. In this task, they judged whether the person presented in the context was more likely to be causing (3), suffering

[2] In the pre-registration, we mistakenly stated "Each context appeared *twice* with each of the…". The correct proportion of assignments is reported in the text.

For Peer Review Only

from (-3), or neither causing nor suffering from (0) danger (for ease of interpretation and consistency with the contextual threat ratings, these scores were linearly transformed to a 1–7 scale by adding 4 to each response prior to analysis and plotting). This block consisted of 16 trials, one for each context stimulus, presented with one of the 8 untrustworthy and the 8 neutral faces.



Figure 1. Examples of ambiguous context stimuli presented alone (left), featuring a neutral face (center), or featuring an untrustworthy face (right).

For Peer Review Only

Results

As pre-registered for all the experiments, data was analyzed using the 'ezANOVA' function (Lawrence, 2016) in R (version 4.4.2) and R basic functions. In case of three-level factors or significant interactions, we conducted pairwise comparisons using the 'emmeans' function (Lenth, 2022) and adjusted the comparisons with the "Holm" method (see Supplementary Materials – Appendix A for analyses conducted using linear mixed models showing unaltered results). See Figure 2 for a graphical representation of all experimental results.

Contextual threat. A repeated measures ANOVA revealed the main effect of face type, $F(1.79, 261.80) = 29.41$, $p < 0.001$, $\eta^2_p = 0.16$ 90% CI [0.10, 0.23] (corrected for violation of sphericity assumption, $\varepsilon = .89$). A pairwise comparison demonstrated that contexts with untrustworthy faces were judged as more threatening ($M = 4.50$, $SD = 1.73$) than those with neutral faces ($M = 4.28$, $SD = 1.72$), $t(292) = 3.98$, $p < 0.001$, $d = 0.23$, 95% CI [0.12, 0.35], and of those with no faces ($M = 4.05$, $SD = 1.80$), $t(292) = 8.09$, $p < 0.001$, $d = 0.47$, 95% CI [0.35, 0.59] (see Figure 2, A panel). Also, contexts with neutral faces were judged as more threatening than those with no faces, $t(292) = 4.12$, $p < 0.001$, $d = 0.24$, 95% CI [0.12, 0.36].

Target role. A paired sample *t*-test showed that targets with untrustworthy faces (M = 5.01, SD = 1.38) were perceived as being more likely to cause danger than neutral faces (M = 3.80, SD = 1.46), $t(146) = 18.52$, $p < .001$, $d = 1.53$, 95% CI [1.29, 1.77] (see Figure 2, B panel).

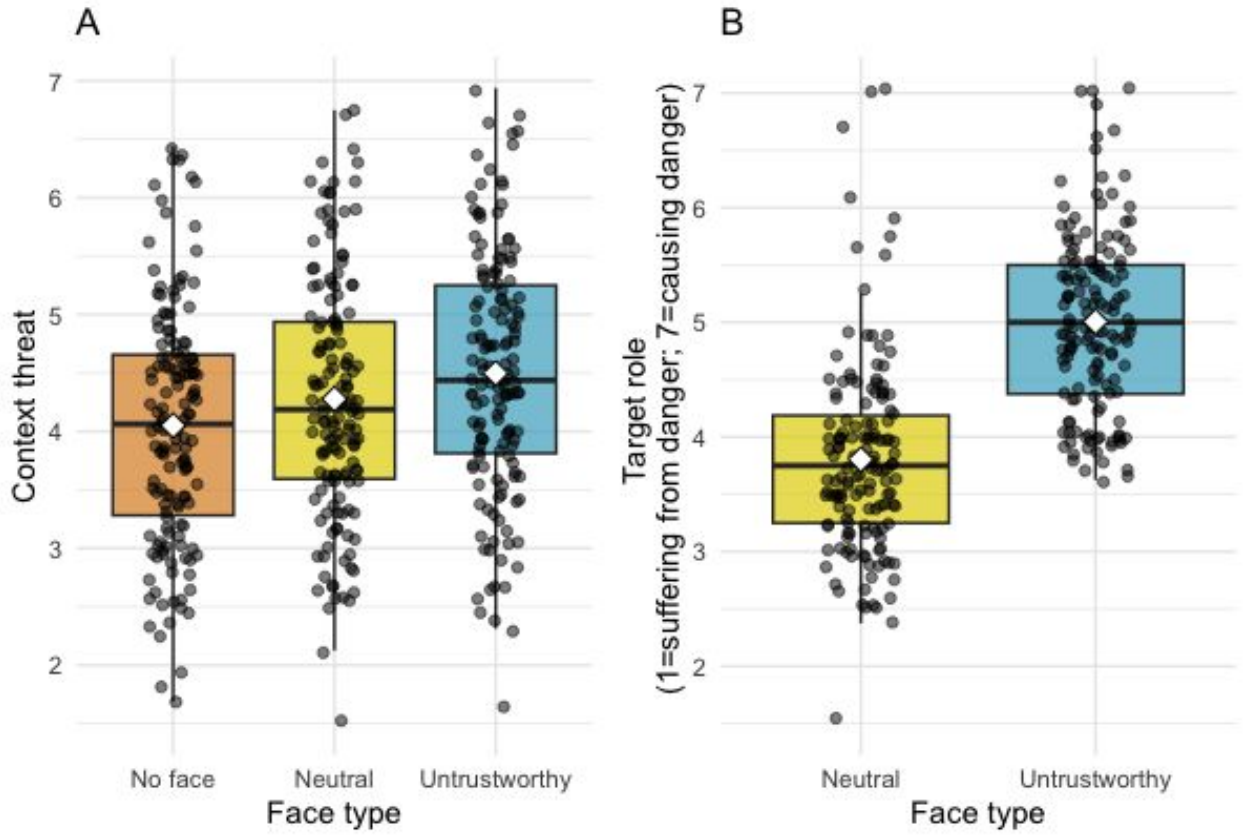


Figure 2. Boxplots with participant-level observations of A) context threat judgements and B) attribution of dangerous intentions depending on face type. The white diamond represents the mean point. Role attribution ratings are displayed on the transformed 1–7 scale (original response scale: −3 to +3).

Association between target role and contextual threat. We conducted a linear mixed-effects analysis. Contextual threat ratings were entered as the dependent variable, with face condition (neutral vs. untrustworthy) and role attribution ratings as predictors. Because face condition was manipulated within participants, participant was included as a random intercept to account for the non-independence of repeated observations. The model was estimated using restricted maximum likelihood (REML), and significance tests were based on Satterthwaite-adjusted degrees of freedom. The linear mixed-effects model revealed that role attribution

significantly predicted contextual threat ratings, $b = 0.31$, $SE = 0.05$, $t(222.48) = 6.05$, $p < .001$. Thus, participants who were more likely to attribute danger to the target also perceived the surrounding context as more threatening, even after accounting for differences between face conditions. The effect of face condition was also significant, although small in magnitude, $b = 0.08$, $SE = 0.04$, $t(185.14) = 1.98$, $p = .049$. This indicates that contextual threat ratings differed slightly between the neutral and untrustworthy face conditions after controlling for role attribution.

Discussion

Evaluations of contexts were influenced by the embedded faces, demonstrating that face-context integration operates in the reverse fashion – from person untrustworthiness to contextual threat. The mere presence of a face made the context more threatening compared to when no face was presented. Moreover, contexts containing untrustworthy (vs. neutral) faces were rated as more threatening. Untrustworthy faces were also more frequently assigned the role of the individual responsible for the danger, and this attribution was associated positively with an increase in perceived contextual threat. Together, these results suggest that facial cues shape environmental appraisal not only through associative links but also through relational inferences about agency and threat causation.

However, in Experiment 1, participants were instructed to imagine themselves in the depicted scenarios when judging contextual threat. It remains unclear whether participants consistently adopted this self-perspective or whether, in some trials, they instead judged to what extent the target person might feel threatened. Even if participants adhered to instructions, this ambiguity leaves open whether the observed face-to-context effects reflect a personal sense of threat or an inferred evaluation of the target's experience. Clarifying this distinction is critical for testing whether face–context integration reflects a simple association between facial and environmental valence or a more complex relational reasoning process that depends on the perceiver's stance toward the scene. To address this issue, Experiment 2 manipulated participants' perspective during contextual threat judgments.

For Peer Review Only

Experiment 2

Experiment 2 examined whether the effects observed in Experiment 1 depend on the perceiver's judgment perspective. An important implication of the relational account is that the meaning of a face-context configuration depends not only on the cues present in the scene but also on the observer's perspective. Relational reasoning requires identifying the role each person plays within the situation. For example, who represents the source of danger and who is the potential recipient of harm. Consequently, the same face–context combination may give rise to different evaluations depending on whether observers interpret the situation from their own perspective or from the target's perspective. If an untrustworthy individual is construed as a potential threat to oneself, contextual threat should increase. By contrast, when observers adopt the target's perspective, the same individual should be no longer interpreted as a threat to the self, weakening the face-to-context influence.

Here, we investigated whether the increase in perceived contextual threat associated with an untrustworthy face differs when participants evaluate threat for themselves versus for the target person depicted in the scene. Participants viewed threatening contexts containing either neutral or untrustworthy faces. Perspective was manipulated between participants: one group rated how threatening the scene felt to them (self-condition), while the other group rated how threatening the scene would be for the target person (target condition). From a relational reasoning standpoint, contextual threat should rise only when the perceiver interprets the face as relevant to their own safety – that is, when they construe the face as a potential source of danger. Thus, we predicted a significant interaction between perspective and facial trustworthiness, such that untrustworthy faces would increase contextual threat only in the self-condition. Finally, we again examined role attributions. We expected untrustworthy faces would be more often categorized as causing danger, whereas neutral faces would be seen as experiencing it. We also tested whether this pattern varied depending on the adopted perspective.

Method

For Peer Review Only

We employed a 2 (face type: neutral vs. untrustworthy faces) x 2 (judgment perspective: self vs. target) design, manipulating the second factor between participants.

Sample Size Determination and Participants

The critical effect was the 2 x 2 interaction effect. We aimed for an average effect (Cohen's d = 0.50). Setting α to 0.05, and power = 0.95, Superpower suggested an overall sample of 208 participants. We slightly oversampled and targeted a total of 210 participants on Prolific Academic. We pre-registered a sequential analysis with a single interim analysis at the midpoint of the sample using the Pocock boundary to set the alpha level. Since the critical test had $p < 0.0294$ when looking at the data of 105 participants (49 women, 54 men, 2 other genders; $M_{age} = 30.24$; $SD_{age} = 8.70$), we stopped the data collection.

Stimuli and Procedure

The experiment was created in Millisecond's Inquisit. After giving consent to participate and providing demographics, participants received the instructions about the stimuli they would see on the screen varying according to the perspective condition they were assigned to. In the Self condition, participants were to indicate to what extent *they* would feel threatened in each scenario. In the Target condition, participants were to indicate the extent to which *the person presented at the center of the screen* would feel threatened in each scenario (see SM for the full instructions). They answered on a scale from 1 (not at all threatened) to 7 (extremely threatened). They were presented with 16 context stimuli. Each context appeared three times, for a total of 48 trials. In 24 trials, 8 context stimuli were paired with one of the 8 untrustworthy faces (repeated three times); in the remaining 24 trials, 8 context stimuli were paired with one of the 8 neutral faces (repeated three times). In a separate block, participants attributed a role to each face presented in the threatening context as in Experiment 1.

Results

Contextual threat. We conducted a 2x2 mixed ANOVA (see Supplementary Materials – Appendix A for analyses conducted using linear mixed models showing unaltered results). We found a main effect of perspective, $F(1, 103) = 12.81$, $p < 0.001$, $\eta_p^2 = 0.11$ 90% CI [0.03, 0.21], meaning

For Peer Review Only

that when using the self-perspective (M = 4.66, SD = 1.70) the context was perceived as more threatening than when using the target perspective (M = 3.94, SD = 1.66), and a main effect of face, $F(1, 103) = 28.11$, $p < 0.001$, $\eta_p^2 = 0.21$ 90% CI [0.11, .32], such that untrustworthy faces (M = 4.51, SD = 1.71) made contexts more threatening than neutral faces (M = 4.21, SD = 1.73). These were qualified by a significant interaction between them, $F(1, 103) = 17.69$, $p < 0.001$, $\eta_p^2 = 0.15$, 90% CI [0.05, 0.25]. Simple effects comparisons showed that untrustworthy faces (M = 4.91, SD = 1.63) made contexts more threatening than neutral faces (M = 4.41, SD = 1.74) in the self-perspective, $t(103) = 6.76$, $p < 0.001$, d = 1.33, 95%CI [0.90, 1.76], but the difference was not significant in the target perspective (untrustworthy faces M = 3.95, SD = 1.65; neutral faces M = 3.93, SD = 1.68), $t(103) = 0.23$, $p = 0.822$, d = 0.04, 95% CI [-0.34, 0.43] (see Figure 3, A panel).

Target role. We conducted a 2x2 mixed ANOVA. We found a main effect of perspective, $F(1, 103) = 14.59$, $p < 0.001$, $\eta_p^2 = .12$ 90% CI [0.04, 0.22], meaning that when using the self-perspective (M = 1.45, SD = 1.52) the target face was seen as more likely to cause danger than when using the target perspective (M = 0.98, SD = 1.48), and a main effect of face, $F(1, 103) = 165.48$, $p < 0.001$, $\eta_p^2 = .62$, 90% CI [0.52, 0.68], such that untrustworthy faces (M =1.83, SD = 1.41) were perceived as more dangerous agents than neutral faces (M = 0.68, SD = 1.40). There was no interaction between the two, $F(1, 103) = 1.03$, $p = 0.31$, $\eta_p^2 = 0.01$, 90% CI [0.00, 0.06] (see Figure 3, B panel).

For Peer Review Only

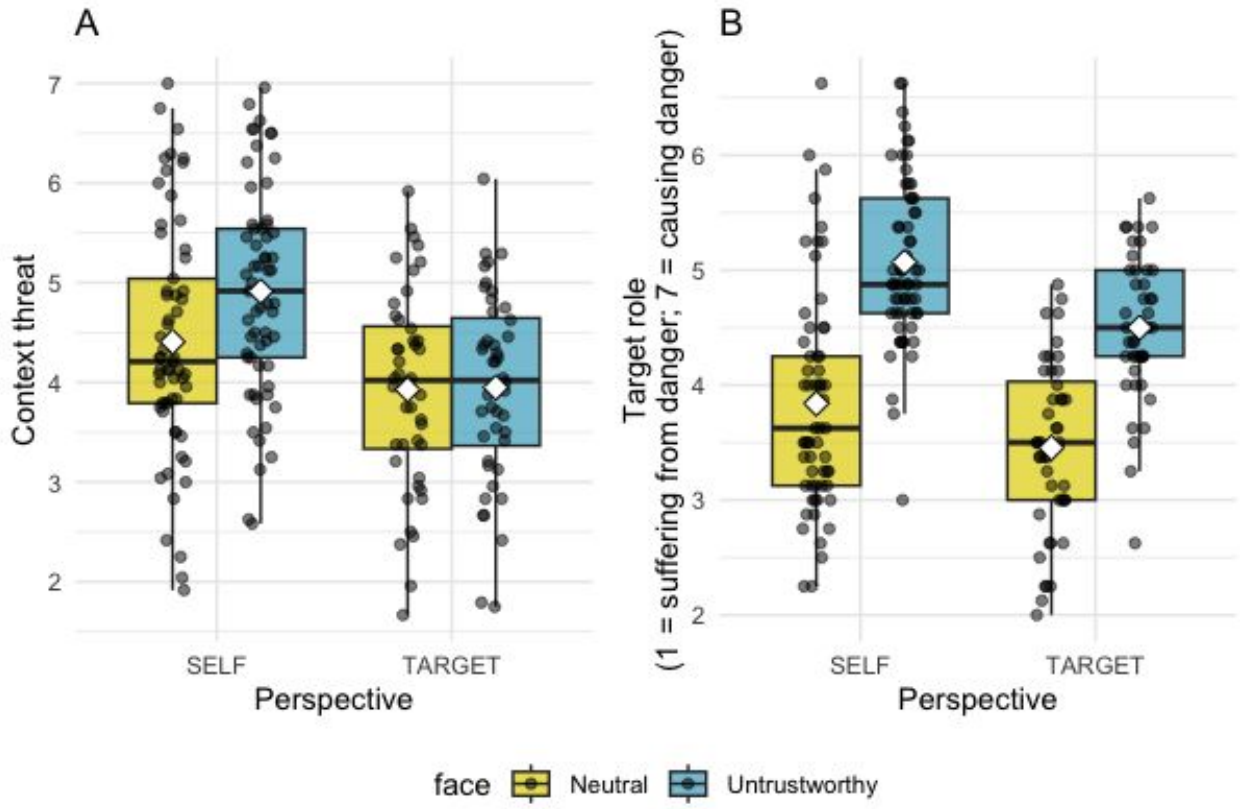


Figure 3. Boxplots with participant-level observations of A) context threat judgements and B) attribution of dangerous intentions depending on the perspective and the face type. The white diamond represents the mean point. Role attribution ratings are displayed on the transformed 1–7 scale (original response scale: −3 to +3).

Association between target role and contextual threat. We conducted a linear mixed-effects model predicting contextual threat ratings from face condition (neutral vs. untrustworthy), perspective (self vs. target), role attribution ratings, and their theoretically motivated interactions (i.e., perspective × role attribution ratings and perspective × face condition). Participant was included as a random intercept to account for the non-independence of repeated observations within participants. The model was estimated using restricted maximum likelihood (REML), with Satterthwaite-adjusted degrees of freedom for significance testing. Role attribution significantly predicted contextual threat ratings, $b = 0.21$, $SE = 0.06$, $t(127.15) = 3.71$, $p < .001$, indicating that participants who attributed greater responsibility for danger to the target also perceived the surrounding context as more threatening. The main effect of face condition was not significant, $b = -0.003$, $SE = 0.04$, $t(113.58) = -0.07$, $p = .947$. A significant effect of perspective emerged, $b = -0.57$, $SE = 0.25$, $t(159.45) =$

−2.24, p = .027. Importantly, the association between role attribution and contextual threat was moderated by perspective, as indicated by a significant role attribution × perspective interaction, b = 0.21, SE = 0.06, t(127.15) = 3.69, p < .001. Follow-up analyses showed that role attribution was a significant predictor of contextual threat in the self-perspective condition, b = 0.42, SE = 0.07, 95% CI [0.27, 0.57], whereas role attribution was unrelated to contextual threat in the target-perspective condition, b = 0.001, SE = 0.09, 95% CI [−0.17, 0.17]. Finally, the condition × perspective interaction was not significant, b = 0.007, SE = 0.04, t(113.58) = 0.16, p = .875.

Discussion

Experiment 2 demonstrated that the perceiver's perspective during evaluation critically modulates face–context integration. The influence of facial trustworthiness on contextual threat emerged only when participants judged the situation from their own standpoint. When asked to adopt the target's perspective, this effect disappeared. Regarding role attribution, participants were more likely to attribute danger to untrustworthy faces than to neutral ones. Although this tendency was descriptively stronger in the self-perspective, the interaction was not significant. Importantly, the association between contextual threat and role attribution was positive and reliable only in the self-perspective. Thus, increases in perceived contextual threat co-occurred with stronger inferences that the face was the source of danger, but only when participants evaluated the scene as something happening to them.

Taken together, these results support a relational account. Contextual threat does not rise merely because an untrustworthy-looking face is present, but because perceivers construct a relational meaning that links the face to the threatening affordances of the environment. When adopting a self-perspective, participants could interpret the target face as a dangerous agent responsible for the threat in the environment. In contrast, adopting the target's perspective disrupted this inference, likely because identifying with the untrustworthy target reduced the sense of threat – "if I am the source of danger, I have nothing to fear".

For Peer Review Only

Experiments 1 and 2 provide converging evidence in support of a relational approach to face–context integration. According to this framework, the influence on contextual threat perception should be specific to domains that afford inferences about an individual's potentially dangerous intentions, such as trustworthiness, and should not extend to other negative dispositional traits that lack this implication. To test this domain specificity, we conducted Experiment 3.

Experiment 3

This experiment aimed to examine whether the face-to-context threat effect is specific to a trait that facilitates threat inferences through the attribution of a dangerous role to the face within the context. If this mechanism is indeed trait-specific, then the effects observed in Experiments 1 and 2 should not generalize to other social traits, such as facial competence, for which threat-related inferences about the role of the target are less intuitive or direct. Participants evaluated the level of threat conveyed by contexts containing embedded faces. Prior to this task, each face underwent a conditioning phase, during which it was associated with behaviors reflecting one of four combinations of moral and competence domains: moral, immoral, competent, or incompetent (see Brambilla et al., 2021). This manipulation allowed us to orthogonally vary trait valence and domain, distinguishing between general negativity and physical threat-specific relational meaning. If the observed integration effects merely reflect a valence-based bias, both immoral and incompetent faces should similarly enhance perceived contextual threat. However, if the effect arises from relational inferences about potential danger, only immoral (i.e., untrustworthy) faces should increase contextual threat relative to moral faces. Conversely, within the domain of competence, no comparable difference was expected. Accordingly, we predicted that untrustworthy faces would amplify contextual threat more than incompetent faces, revealing that face–context integration is trait-specific and, in the case of trustworthiness, functionally grounded in threat attribution.

Method

We employed a 2 (target domain: morality vs. competence) x 2 (target valence: positive vs. negative) within-subjects design.

For Peer Review Only

Sample Size Determination and Participants

Of interest for the present study was the 2 x 2 interaction effect. We conducted a Monte-Carlo simulation using the superpower() function in R. Setting α to 0.05, collecting 200 participants will give us 84% of power to detect an expected interaction effect $\eta_p^2 = 0.047$. Thus, our target was a total of 230 Prolific users. We preregistered a sequential analysis to conduct a single interim analysis at half of the total sample. Using the Pocock boundary to set the alpha level, we found the critical test to have $p < .0294$, so we stopped at 115 participants (69 women, 44 men, 2 other genders, M_{age} = 33.21; SD_{age} = 9.34).

Stimuli

Faces. We employed 16 neutral face identities (M = 3.41 [3.25; 3.59]; SD = 0.11) taken from the Chicago Face Database (Ma et al., 2015).

Contexts. The contexts were the same used in the previous experiments.

Behaviors. For the manipulation of the social domain, we created 16 descriptions of behaviors in the domains of morality and competence. We pretested and selected 4 moral (e.g., "volunteering at the red cross'", M = [1.79; 2.49], SD = [.63; 1.04]), 4 immoral (e.g., "teaching children to avoid making friends with non-Italian people", M = [-1.88; -2.65], SD = [.81; 1.25]), 4 competent (e.g., "speak three languages fluently", M = [2.00; 2.65], SD = [.67; 1.18]) and 4 incompetent (e.g., "repeatedly losing their wallet", M = [-2.02; -2.56], SD = [.85; 1.24]) behaviors. The full list is reported in the Supplementary Materials – Appendix B.

Procedure

The experiment comprises 2 different blocks (i.e., a positive block and a negative block)[3], each composed of a learning phase, a memory test, and an evaluation phase. The blocks' order was randomized between participants. After giving consent to participate and providing demographics,

[3] We considered two alternative block structures (blocks organized by valence versus blocks organized by trait domain) before designing the study. We ultimately chose to organize blocks by valence to avoid making the positive-negative distinction overly salient to participants. If blocks had been organized by domain, participants might have inferred that the critical distinction concerned the evaluative polarity of the traits within each domain and could have strategically recoded the tasks in terms of valence.

For Peer Review Only

participants underwent a learning task. They were presented with 8 pairs of faces and behaviors and instructed to pay close attention to both. In the positive block, each face was paired with either moral or competent behavior. Pairs were randomized and presented 4 times each, for a total of 32 trials. To facilitate learning, each behavior was accompanied by a tag signaling whether the behavior was either moral or competent. At the end of the learning phase, participants performed a memory task: we presented again the facial stimuli employed in the learning task, asking participants to identify whether the face had previously been associated with a moral vs. competent behavior. Next, they judged the perceived threat of the face-context pairs as in Experiment 2. Each face was presented twice within one of the contexts, for a total of 16 trials. The negative block was identical, except for (a) the behaviors employed in the learning phase, which were either immoral or incompetent, and (b) the face identities, which varied from those used in the positive block.

Results

We conducted a 2x2 ANOVA with face valence and domain effects (see Supplementary Materials – Appendix A for analyses conducted using linear mixed models showing unaltered results). We found a main effect of face valence, $F(1, 114) = 32.26$, $p < 0.001$, $\eta_p^2 = 0.22$, 90% CI [0.12, 0.32], meaning that negative faces ($M = 4.85$, $SD = 1.69$) made contexts more threatening than positive faces ($M = 4.49$, $SD = 1.67$), and a main effect of domain, $F(1, 114) = 5.74$, $p = 0.02$, $\eta_p^2 = 0.05$, 90% CI [0.004, 0.12], showing that faces qualified for trustworthiness ($M = 4.74$, $SD = 1.73$) made contexts more threatening than those qualified for competence ($M = 4.60$, $SD = 1.75$). The interaction was significant, $F(1, 114) = 14.30$, $p < 0.001$, $\eta_p^2 = 0.11$, 90% CI [0.03, 0.20]. A simple effect analysis showed that when faces were negative, contextual threat was higher for untrustworthy ($M = 5.03$, $SD = 1.66$) than incompetent faces ($M = 4.66$, $SD = 1.71$), $t(228) = 4.35$, $p < 0.001$, $d = 0.58$, 95% CI [0.31, 0.84], but when looking at positive faces, there was no difference between trustworthy ($M = 4.45$, $SD = 1.75$) and competent faces ($M = 4.53$, $SD = 1.78$), $t(228) = 0.92$, $p = 0.364$, $d = 0.12$, 95% CI [-0.14, 0.34]. Moreover, within the morality domain, untrustworthy faces made contexts more threatening than trustworthy faces, $t(228) = 6.72$, $p < 0.001$, $d = 0.89$, 95% CI

For Peer Review Only

[0.62, 1.16], but within the competence domain there was no difference, $t(228) = 1.48$, $p = 0.140$, $d = 0.20$, 95% CI [-0.06, 0.46] (see Figure 4).[4]

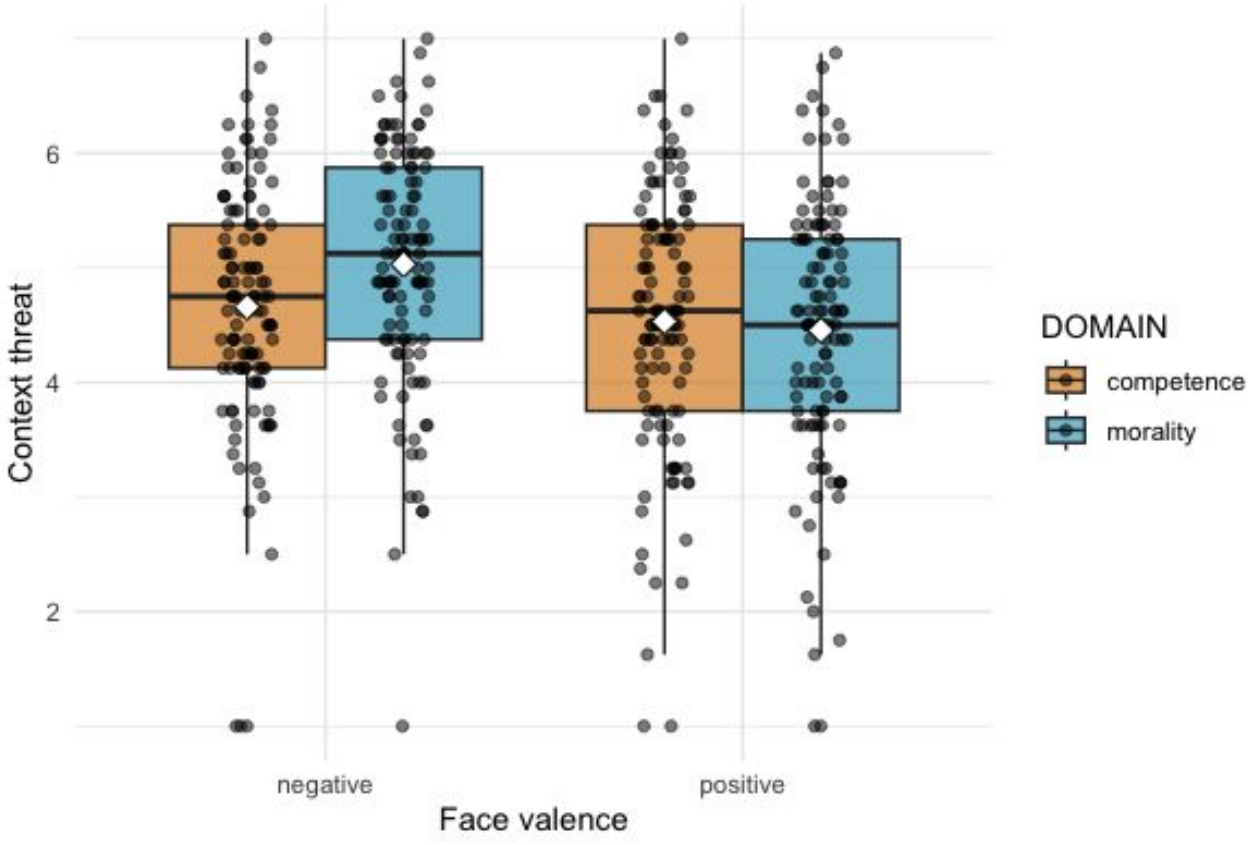


Figure 4. Boxplots with participant-level observations of context threat judgements depending on domain and face valence. The white diamond represents the mean point.

Discussion

We found that perceptions of contextual threat were more strongly influenced by facial trustworthiness than by competence: in particular, it increased when the face was untrustworthy rather than trustworthy. This finding contradicts a general negativity bias for any social dimension but supports a preferential trust-threat association that can be explained by a relational account. To assess contextual threat, participants should make meaningful inferences about the dangerous role of the

[4] An anonymous reviewer noticed that because two of the behavioural descriptions used in the immoral condition referred to driving violations, they may have been especially relevant to some of the road scenes used in the contextual threat task. To assess whether this overlap influenced the results, we repeated the main analysis after excluding the face identities associated with these behavioural descriptions. The pattern of results remained unchanged, with the same main effects and Face Valence × Domain interaction remaining statistically significant.

target in the context. When targets were untrustworthy, such inference was easier, hence the increase in contextual threat. Instead, when targets were incompetent, facial features did not provide useful information for making inferences about contextual threat and were thus less effective. This interpretation should nevertheless be considered in light of the design approach adopted in Experiment 3. Although we organized blocks by valence to avoid making the mere positive-negative distinction overly salient across dimensions (i.e., trustworthiness vs. competence), replicating the asymmetric effect with domain-based blocks would provide further support for the present findings.

General discussion

Across three experiments, we found that perceived face trustworthiness systematically shapes evaluations of environmental threat. Experiment 1 showed that the mere presence of a face (vs. no face) in a scene increased perceived contextual threat, with untrustworthy faces amplifying this effect more than neutral ones. This heightened threat was associated with perceiving untrustworthy faces as dangerous. Experiment 2 revealed that the reported effect depends on the perspective adopted in the judgment. Untrustworthy faces increased contextual threat only when participants evaluated the scene from their own perspective; when imagining the scene from the target's perspective, this difference disappeared. Role attributions remained biased, with untrustworthy faces seen as more dangerous agents, but their link to contextual threat was evident only in the self-perspective condition. Experiment 3 showed that this 'face-to-context' effect is trait-specific: untrustworthy (vs. trustworthy) faces increased perceived contextual threat, whereas incompetent (vs. competent) faces did not. Together, these findings indicate that trustworthiness uniquely informs inferences about potential danger and that such inferences can reshape the perceived social meaning of the broader environment.

Our work advances the literature on face–context integration, which has largely emphasized how threatening contexts can reduce trustworthiness ascribed to target individuals (Brambilla et al., 2018, 2021, 2024; Mattavelli et al., 2022), often via inferences about the individual's role in the scene (Mattavelli et al., 2023a). We showed that the reverse pathway is also viable: facial untrustworthiness

For Peer Review Only

can heighten the perceived threat of an otherwise ambiguous environment. This reversal is theoretically significant, as most existing models implicitly assume that environments shape impressions of people. Instead, our data suggests that person and context jointly define each other's meaning, with evaluative information flowing bidirectionally to maintain coherence within a social scene. In this sense, social perception operates as a system in which the interpretation of people and places is dynamically adjusted to yield consistent evaluations.

From a theoretical stance, our findings align with a propositional view of the face-context integration phenomenon (see also Mattavelli et al., 2023a, 2023b). Borrowed from conditioning research (De Houwer, 2014, 2018, 2020), this propositional view suggests that the meaning of one element depends on its inferred relation to another. Here, we extend it by showing that the causal link can run from individuals to their surrounding environment. Although the present studies do not directly manipulate relational reasoning, three lines of evidence are consistent with this interpretation. First, the effect emerges only when observers adopt their own perspective in evaluating contextual threat, consistent with the idea that perceivers represent themselves as potential recipients of harm. Second, the strength of the effect correlates with perceiving the target as a threatening agent in the scene. Although correlational in nature, this finding is more readily explained by relational role attribution than by a simple associative transfer of affect. Third, the absence of an effect for competence confirms that this process hinges on trustworthiness, a trait functionally linked to threat detection (Brambilla & Leach, 2014; Cosmides & Tooby, 1992). Together, these findings provide converging evidence that the observed face-to-context effects are compatible with a relational account of face–context integration. Future research should directly manipulate relational structure (e.g., causal roles or relational cues) to test this mechanism more explicitly.

The results of Experiment 3 also carry important implications for understanding the mechanism underlying face–context integration and for interpreting the role of facial trustworthiness. By using behavioral information to signal trustworthiness rather than relying only on spontaneous inferences from facial appearance, we demonstrate that the effect is not limited to static

For Peer Review Only

morphological cues or to confounding factors that often accompany physical untrustworthiness (e.g., facial stereotypes or aesthetic preferences). In fact, dispositional information inferred from immoral behaviors of target faces similarly altered perceptions of an otherwise neutral environment. This suggests that face–context integration operates at a representational level beyond mere visual perception, encompassing conceptual inferences about moral character and their implications for situational threat appraisal. Accordingly, the trust–threat link reflects an inference-driven process that flexibly integrates trait information with contextual evaluation.

This work is not without limitations. First, all facial stimuli depicted male identities. Consequently, the present findings cannot be assumed to generalize to female faces. This limitation is theoretically relevant because facial trustworthiness is naturally embedded within broader gendered appearance cues, and male and female faces may differ in how trustworthiness and threat-related information are integrated (Oh et al., 2020). Prior research on face-context integration showed that contextual threat has a comparable influence on male and female (un)trustworthiness (Mattavelli et al., 2023a, Experiment 1c); yet, it remains unclear whether target gender plays a role in the reverse relationship, namely, in the influence of person trustworthiness on contextual threat.

Second, the stimuli in all experiments were static and decontextualized in time. Real-world encounters unfold dynamically, with facial cues, behaviors, and environmental changes occurring in rapid succession. Future work should examine whether similar effects emerge in dynamic or immersive scenes. Third, in Experiments 1 and 2, we selected faces based on normative trustworthiness ratings. Although trustworthiness was the focal facial dimension of interest, naturally occurring facial impressions are correlated with other characteristics, such as dominance, aggressiveness, masculinity, and emotional expressions. Consequently, we cannot completely rule out the possibility that these traits also contributed to the observed effects. Importantly, however, our aim was not to isolate trustworthiness from all other facial characteristics, but rather to examine whether facial information that naturally conveys potential threat influences evaluations of contextual threat. Future research could further disentangle the unique contributions of trustworthiness and other

For Peer Review Only

correlated facial dimensions by independently manipulating these characteristics using computationally generated faces. Fourth, our measures of contextual threat relied exclusively on self-report, which may not capture fast, automatic components of threat appraisal. Importantly, both these aspects could make participants more sensitive to demand characteristics (Corneille & Lush, 2023). That is, knowing they are being asked to evaluate threat in relation to faces might lead participants to infer the "expected" pattern and adjust their responses accordingly. Future research could address these limitations by employing dynamic or immersive stimuli, as well as implicit or physiological measures of threat, reducing the potential for responses to be shaped by perceived experimental expectations. Such approaches would clarify whether the observed face-to-context effects genuinely reflect inference-driven integration rather than participant expectations.

Conclusions

Across three experiments, we show that perceived trustworthiness systematically shapes evaluations of contextual threat. Untrustworthy faces increase the perceived danger of otherwise ambiguous environments. Highlighting the role of relational reasoning in qualifying the link between facial trustworthiness and contextual threat, we found that (i) this effect correlated with role attribution, (ii) emerged only when observers adopted their own perspective, and (iii) was specific for trustworthiness manipulations of the target face. Together, these findings advance a relational account of social perception, where person information dynamically reshapes how environments are appraised, and vice versa.

Conflict of Interest statement

The authors report there are no competing interests to declare

Funding statement

This research received no external funding.

For Peer Review Only

References

Aguado, L., Parkington, K. B., Dieguez-Risco, T., Hinojosa, J. A., & Itier, R. J. (2019). Joint modulation of facial expression processing by contextual congruency and task demands. *Brain Sciences*, *9*(5), 116. https://doi.org/10.3390/brainsci9050116

Aviezer, H., Bentin, S., Dudarev, V., & Hassin, R. R. (2011). The automaticity of emotional face-context integration. *Emotion*, *11*(6), 1406-1414. https://doi.org/10.1037/a0023578

Barrett, L. F., & Kensinger, E. A. (2010). Context is routinely encoded during emotion perception. *Psychological Science*, *21*(4), 595-599. https://doi.org/10.1177/0956797610363547

Brambilla, M., & Leach, C. W. (2014). On the importance of being moral: The distinctive role of morality in social judgment. *Social Cognition*, *32*(4), 397-408. https://doi.org/10.1521/soco.2014.32.4.397

Brambilla, M., Biella, M., & Freeman, J. B. (2018). The influence of visual context on the evaluation of facial trustworthiness. *Journal of Experimental Social Psychology*, *78*, 34-42. https://doi.org/10.1016/j.jesp.2018.04.011

Brambilla, M., Masi, M., Mattavelli, S., & Biella, M. (2021). Faces and sounds becoming one: Cross-modal integration of facial and auditory cues in judging trustworthiness. *Social Cognition*, *39*(3), 315-327. https://doi.org/10.1521/soco.2021.39.3.315

Brambilla, M., Mattavelli, S., & Masi, M. (2024). Face–context integration and trustworthiness evaluation. *European Review of Social Psychology*, *35*(2), 378-423. https://doi.org/10.1080/10463283.2024.2315915

Corneille, O., & Lush, P. (2023). Sixty years after Orne's American psychologist article: A conceptual framework for subjective experiences elicited by demand characteristics. *Personality and Social Psychology Review*, *27*(1), 83-101. https://doi.org/10.1177/10888683221104368

For Peer Review Only

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. *The adapted mind: Evolutionary psychology and the generation of culture*, *163*, 163-228.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology*, *10*(2), 230-241. https://doi.org/10.1017/S1138741600006491

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342-353. https://doi.org/10.1111/spc3.12111

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, *13*(3), 1-21. https://doi.org/10.5964/spb.v13i3.28046

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. In *Advances in experimental social psychology* (Vol. 61, pp. 127-183). Academic Press.

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management*, *55*, 62-73. https://doi.org/10.1016/j.tourman.2016.01.013

Fiedler, K., & Unkelbach, C. (2011). Evaluative conditioning depends on higher order encoding processes. *Cognition and Emotion*, *25*(4), 639-656. https://doi.org/10.1080/02699931.2010.513497

Fiske, S. T. (1992). Thinking is for doing: portraits of social cognition from daguerreotype to laserphoto. *Journal of Personality and Social Psychology*, *63*(6), 877 –889. https://doi.org/10.1037/0022-3514.63.6.877

Foo, Y. Z., Sutherland, C. A. M., Burton, N. S., Nakagawa, S., & Rhodes, G. (2022). Accuracy in facial trustworthiness impressions: Kernel of truth or modern physiognomy? A meta-analysis. *Personality and Social Psychology Bulletin*, *48*(11), 1580-1596. https://doi.org/10.1177/01461672211048110

For Peer Review Only

Förderer, S., & Unkelbach, C. (2012). Hating the cute kitten or loving the aggressive pit-bull: EC effects depend on CS–US relations. *Cognition and Emotion*, *26*(3), 534-540. https://doi.org/10.1080/02699931.2011.588687

Hess, U., Dietrich, J., Kafetsios, K., Elkabetz, S., & Hareli, S. (2020). The bidirectional influence of emotion expressions and context: Emotion expressions, situational information and real-world knowledge combine to inform observers' judgments of both the emotion expressions and the situation. *Cognition and Emotion*, *34*(3), 539-552. https://doi.org/10.1080/02699931.2019.1651252

Hess, U., Landmann, H., David, S., & Hareli, S. (2017). The bidirectional relation of emotion perception and social judgments: The effect of witness' emotion expression on perceptions of moral behaviour and vice versa. *Cognition and Emotion*, 1–14. doi: 10.1080/02699931.2017.1388769

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4*(1), https://doi.org/10.1177/2515245920951503

Lawrence, M. A., & Lawrence, M. M. A. (2016). Package 'ez'. *R package version*, *4*(0), 1620-1631. http://github.com/mike-lawrence/ez

Lenth, R. (2022). emmeans: Estimated marginal means, aka least-squares means. R package version 1.7. 2.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122-1135. https://doi.org/10.3758/s13428-014-0532-5

Masuda T, Ellsworth PC, Mesquita B, Leu J, Tanida S, Van de Veerdonk E. 2008. Placing the face in context: cultural differences in the perception of facial emotion. *Journal of Personality and Social Psychology. 94*:365–381. https://doi.org/10.1037/0022-3514.94.3.365

For Peer Review Only

Mattavelli, S., Fiamberti, G. C., Masi, M., & Brambilla, M. (2023b). The "Happy Face Killer" in the eyes of the beholder: Relational encoding of facial emotions in context influences trustworthiness attributions. *Journal of Experimental Social Psychology*, *109*, 104517. https://doi.org/10.1016/j.jesp.2023.104517

Mattavelli, S., Masi, M., & Brambilla, M. (2022). Untrusted under threat: On the superior bond between trustworthiness and threat in face-context integration. *Cognition and Emotion*, *36*(7), 1273-1286. https://doi.org/10.1080/02699931.2022.2103100

Mattavelli, S., Masi, M., & Brambilla, M. (2023a). Not just about faces in context: Face–context relation moderates the impact of contextual threat on facial trustworthiness. *Personality and Social Psychology Bulletin*, *49*(3), 376-390. https://doi.org/10.1177/01461672211065933

Moran, T., & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognition & Emotion*, *27*(4), 743-752. https://doi.org/10.1080/02699931.2012.732040

Oh, D., Dotsch, R., Porter, J., & Todorov, A. (2020). Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General, 149*(2), 323–342. https://doi.org/10.1037/xge0000638

Righart, R., & De Gelder, B. (2008). Recognition of facial expressions is influenced by emotional scene gist. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(3), 264-272. https://doi.org/10.3758/CABN.8.3.264

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*(3), 409–426. https://doi.org/10.1037/a0031050

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*(6), 813-833. https://doi.org/10.1521/soco.2009.27.6.813

For Peer Review Only

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455-460. https://doi.org/10.1016/j.tics.2008.10.001

Wieser, M. J., & Brosch, T. (2012). Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in Psychology*, *3*, 471. https://doi.org/10.3389/fpsyg.2012.00471

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592-598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*(8), 1325-1331. https://doi.org/10.1177/0956797615590992

For Peer Review Only

For Peer Review Only

Supplementary Materials

Appendix A. Analyses with mixed models

For all the studies we determined random effects in all models using the *buildmer* package in R (Voeten, 2020).

Experiment 1

Contextual threat. The final random structure included the slope by participants and the intercepts by participants and facial stimuli. The effect of face type was significant, $F(2, 145.97) = 39.71$, $p < .001$. Pairwise comparisons (Bonferroni-corrected) revealed that responses were contextual threat was significantly higher in the untrustworthy face condition than in the neutral face condition, $t(146) = 6.81$, $p < .001$, and in the neutral face condition compared to the no face condition, $t(146) = 4.16$, $p < .001$. Contextual threat was also significantly higher in the untrustworthy face condition compared to the no face condition, $t(146) = 8.34$, $p < .001$.

Role attribution. The final random structure included both the slope and the intercept by participants. The effect of face type was significant, $b = -0.60$, $SE = 0.03$, $t(146) = -18.52$, $p < .001$, indicating that untrustworthy faces were perceived as more likely to cause danger than neutral faces.

Experiment 2

Contextual threat. We centered the two independent variables (perspective: self = 0.5; target = -0.5; face: neutral = 0.5; untrustworthy = -0.5). The final model included the intercepts by participants and facial stimuli as well as the slopes for the effect of face by participants and face stimuli. We found a main effect of perspective, $b = 0.72$, $SE = 0.20$, $t(99.07) = 3.66$, $p < .001$, indicating that the context was perceived as more threatening in the self than in the target perspective. We also found a main effect of face, $b = -0.25$, $SE = 0.05$, $t(45.21) = -4.61$, $p < .001$, with untrustworthy faces making contexts more threatening than neutral faces. A significant interaction, $b = -0.50$, $SE = 0.11$, $t(100.35) = -4.74$, $p < .001$, revealed that untrustworthy faces made contexts stimuli more threatening than neutral faces in the self-perspective, $b = 0.50$, $SE =$

For Peer Review Only

0.08, $t(52.80) = 6.58$, $p < .001$, but not significant in the target perspective, $b = 0.004$, $SE = 0.08$, $t(39.90) = 0.05$, $p = .958$.

Role attribution. We centered the two independent variables (perspective: self = 0.5; target = -0.5; face: neutral = 0.5; untrustworthy = -0.5). The final model included the slopes for the effect of face and perspective by participants. We found a main effect of perspective, $b = 0.48$, $SE = 0.12$, $t(102.05) = 4.04$, $p < .001$, meaning that when using the self-perspective the target face was seen as more likely to cause danger than when using the target perspective. We also found a main effect of face, $b = -1.14$, $SE = 0.09$, $t(103.00) = -12.53$, $p < .001$, with untrustworthy faces perceived as more dangerous than neutral faces. The interaction was not significant, $b = 0.18$, $SE = 0.18$, $t(103) = -1.01$, $p = .313$.

Experiment 3

Contextual threat. We centered the two independent variables (domain: morality = -0.5; competence = 0.5; valence: positive = 0.5; negative = -0.5). The final model included the intercepts by participants and facial stimuli as well as the slopes for the interaction between valence and domain by participants and the effect of domain by visual stimuli. We found a main effect of face valence, $b = -0.35$, $SE = 0.06$, $t(113.94) = -5.64$, $p < .001$, indicating that negative faces made contexts more threatening than positive faces. A main effect of domain emerged, $b = -0.15$, $SE = 0.06$, $t(35.03) = -2.37$, $p = .024$, showing that faces qualified for trustworthiness made contexts more threatening than that those qualified for competence. The interaction was significant, $b = 0.49$, $SE = 0.11$, $t(113.75) = 4.30$, $p < .001$. For faces conditioned to be negative, contextual threat was higher for immorality than incompetence, $b = 0.39$, $SE = 0.10$, $t(88) = 3.97$, $p < .001$, while no significant difference emerged for positive faces, $b = -0.09$, $SE = 0.07$, $t(43.2) = -1.34$, $p = .187$.

For Peer Review Only

Appendix B. List of behaviors used in Experiment 3 (translated from Italian)

The person…

Moral

Works in a community for people with disabilities

Volunteers at the Red Cross

Brings used clothes and food to the homeless every week

In his free time, he works as a clown in the pediatric oncology ward

Immoral

Teaches his children to avoid non-Italian people

Never respects speed limits, not even near a school

Doesn't pay rent to the landlord

Drives after spending hours drinking at the pub every Thursday evening

Competent

Speaks three languages fluently

Graduated in medicine with honors

Won a scholarship

Founded a successful start-up

Incompetent

Repeatedly loses his wallet

Was fired from two companies in one year

Submitted a CV full of mistakes

Was rated the worst employee of the year

For Peer Review Only

Appendix C. Facial stimuli normative scores

Table S1 reports the normative scores for the 16 facial stimuli used in Experiments 1-2 on trustworthiness (key manipulation), dominance, masculinity, and anger. The two sets also differed in dominance, with untrustworthy faces being perceived as more dominant than neutral faces, $t(14) = -2.68$, $p = .018$, and in anger, with untrustworthy faces being perceived as angrier, $t(14) = -8.85$, $p < .001$. By contrast, the two sets did not differ in perceived masculinity, $t(14) = -0.37$, $p = .718$.

Table S1. Facial stimuli normative score from Chicago Face Database (Ma et al., 2015)

CFD ID	Study set	Trust	Dominance	Masculinity	Anger
CFD-WM-014	neutral	3.81	3.40	4.87	2.29
CFD-WM-023	neutral	3.81	2.50	4.55	1.76
CFD-WM-024	neutral	3.78	2.28	3.95	1.94
CFD-WM-029	neutral	3.70	2.82	5.21	1.77
CFD-WM-208	neutral	3.59	2.18	3.31	1.68
CFD-WM-214	neutral	3.72	2.48	3.67	1.68
CFD-WM-225	neutral	3.70	3.11	4.61	1.93
CFD-WM-257	neutral	3.92	2.19	4.35	1.33
CFD-WM-010	untrust	2.46	2.80	3.79	3.61
CFD-WM-035	untrust	2.76	2.96	4.48	3.84
CFD-WM-220	untrust	2.48	4.62	5.07	3.74
CFD-WM-222	untrust	2.76	2.57	4.07	2.75
CFD-WM-224	untrust	2.82	4.18	4.75	3.14
CFD-WM-235	untrust	2.57	3.07	4.32	3.00
CFD-WM-243	untrust	2.56	3.88	4.65	3.20
CFD-WM-249	untrust	2.31	3.33	4.15	3.04

Note: each variable was measured with a 1-7 Likert scale

For Peer Review Only