Can people detect the trustworthiness of strangers based on their facial appearance?

Bastian Jaeger\textsuperscript{a, b, *}, Bastiaan Oud\textsuperscript{c}, Tony Williams\textsuperscript{c}, Eva G. Krumhuber\textsuperscript{d}, Ernst Fehr\textsuperscript{c}, Jan B. Engelmann\textsuperscript{e, f, g}

\textsuperscript{a} Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, the Netherlands
\textsuperscript{b} Department of Social Psychology, Tilburg University, the Netherlands
\textsuperscript{c} Department of Economics, University of Zurich, 8006 Zurich, Switzerland
\textsuperscript{d} Department of Experimental Psychology, University College London, United Kingdom
\textsuperscript{e} Amsterdam Brain and Cognition, University of Amsterdam, the Netherlands
\textsuperscript{f} Center for Research in Experimental Economics and Political Decision Making, University of Amsterdam, the Netherlands
\textsuperscript{g} Behavioral and Experimental Economics, The Tinbergen Institute, the Netherlands

\textbf{A R T I C L E  I N F O}

\textbf{Keywords:}
Trust
Trustworthiness
Cooperation
Face perception
Prediction
Accuracy

\textbf{A B S T R A C T}

Although cooperation can lead to mutually beneficial outcomes, cooperative actions only pay off for the individual if others can be trusted to cooperate as well. Identifying trustworthy interaction partners is therefore a central challenge in human social life. How do people navigate this challenge? Prior work suggests that people rely on facial appearance to judge the trustworthiness of strangers. However, the question of whether these judgments are actually accurate remains debated. The present research examines accuracy in trustworthiness detection from faces and three moderators proposed by previous research. We investigate whether people show above-chance accuracy (a) when they make trust decisions and when they provide explicit trustworthiness ratings, (b) when judging male and female counterparts, and (c) when rating cropped images (with non-facial features removed) and uncropped images. Two studies showed that incentivized trust decisions (Study 1, n = 131 university students) and incentivized trustworthiness predictions (Study 2, n = 266 university students) were unrelated to the actual trustworthiness of counterparts. Accuracy was not moderated by stimulus type (cropped vs. uncropped faces) or counterparts’ gender. Overall, these findings suggest that people are unable to detect the trustworthiness of strangers based on their facial appearance, when this is the only information available to them.

\textsuperscript{*} Corresponding author at: Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, the Netherlands.
E-mail address: bxjaeger@gmail.com (B. Jaeger).

https://doi.org/10.1016/j.evolhumbehav.2022.04.004
Received 30 December 2020; Received in revised form 18 November 2021; Accepted 10 April 2022

Please cite this article as: Bastian Jaeger, Evolution and Human Behavior, https://doi.org/10.1016/j.evolhumbehav.2022.04.004
Previous studies have examined the accuracy of trustworthiness impressions in the context of social dilemma games such as the trust game (Berg, Dickhaut, & McCabe, 1995). In this dyadic interaction, a participant (i.e., the trustor) decides whether to send a monetary endowment to another participant (i.e., the trustee). If the endowment is transferred, the money is multiplied and the trustee decides how much to return to the trustor. Trust and reciprocity lead to higher payoffs for both, but trust is risky as trustees face the temptation to keep the transferred money. Bonnefon, Hopfensitz, and De Neys (2013) presented facial photographs of trustees that had either reciprocated or betrayed trust, showing that participants were more likely to transfer money to counterparts that were actually trustworthy. Other studies yielded similar results, leading various authors to conclude that people are able to detect the trustworthiness of counterparts at levels slightly above chance (ca. 55%; Bonnefon, Hopfensitz, & De Neys, 2017; De Neys, Hopfensitz, & Bonnefon, 2013, 2015, 2017; Tognetti, Berticat, Raymond, & Faurie, 2013; Verplaetse, Vanneste, & Braeckman, 2007). Yet, evidence of accurate trustworthiness detection is mixed. Some research did not find empirical support for accuracy when examining trust behavior in social dilemma games (Efferson & Vogt, 2013; Yama- gishi, Tanida, Mashima, Shimosa, & Kanazawa, 2003) or when obtaining explicit ratings of counterparts varying in trustworthiness (Rule, Krendl, Ivcevic, & Ambady, 2013; Zylbersztejn, Babutsidze, & Hanaki, 2020). Moreover, accuracy often depended on extraneous factors, which did not replicate across studies. For example, Tognetti et al. (2013) found above-chance accuracy for male but not female counterparts, when using images that were uncropped and included non-facial features (e.g., hair style). Bonnefon et al. (2013), on the other hand, found higher levels of accuracy for female counterparts, but only with cropped images that occluded all non-facial features.

Various scholars have also criticized the accuracy claim by arguing that the reliability of any facial feature as an indicator of trustworthiness might be easily undermined if individuals exhibit the feature but act selfishly (Efferson & Vogt, 2013; McCullough & Reed, 2016). This could lead to the emergence of imitators who appear trustworthy and garner the benefits of trust without paying the costs of reciprocating it. Furthermore, trustworthiness impressions of the same individual vary substantially across different perceivers (Heman, Sutherland, Flack, & Slepian, 2017) and contexts (Brambilla, Biella, & Freeman, 2018), questioning whether they could be a reliable indicator of any disposition. In sum, the existing evidence for accurate trustworthiness detection from faces is mixed and the topic remains subject to vigorous debate.

1.1. Aims of the present research

We present the results of two studies on the accuracy of trustworthiness impressions that address critical limitations of prior work. First, many prior studies relied on the same set of facial photographs (Bonnefon et al., 2013; De Neys et al., 2013, 2015, 2017) and explicitly selected photographs of trustees that were judged with the highest levels of accuracy in prior investigations (De Neys et al., 2015, 2017). These results do not provide independent accuracy estimates and it is unclear whether findings generalize to other stimulus sets. Here, we provide a stronger test of the generalizability of prior results by examining accuracy using novel samples of participants and stimuli.

Second, past research uncovered several moderators (e.g., above-chance accuracy for female, but not male counterparts), but these moderators did not emerge consistently across studies (De Neys et al., 2013; Tognetti et al., 2013). We examine the robustness of the proposed moderators by testing whether participants show above-chance accuracy (a) when they make trust decisions vs. provide explicit trustworthiness ratings, (b) when they rate cropped images (with non-facial features being removed) vs. uncropped images, and (c) when the trustee is male vs. female.

Third, several scholars have posited that facial appearance is not indicative of actual trustworthiness (Efferson & Vogt, 2013; Todorov, Oliva et al., 2015). Yet, existing studies have exclusively focused on statistical methods that cannot provide evidence for such a null hypothesis. The present research addresses this issue by reporting the results of Bayesian analyses (alongside frequentist statistics), which can quantify evidence in favor of the null hypothesis (Wagenmakers, 2007).

Fourth, unlike many previous studies (Bonnefon et al., 2013; De Neys et al., 2013, 2015, 2017; Rule et al., 2013), we draw upon experimental methods from economics and use only fully incentivized experiments that avoided any deception (Camerer & Mobbs, 2017; Hertwig & Ortmann, 2001; Ortmann & Hertwig, 2002). This approach directly couples participants’ responses to their financial payoffs in both the trust game and the prediction task, thereby motivating participants to correctly identify trustworthiness in their counterparts because this increases their own payouts. Decisions during payout relevant trials were paid out not only to thetrustors, but also to the interaction partners whose faces were shown to maintain the interactive nature of the trust game (e.g., Engemann, Meyer, Ruff, & Fehr, 2019). Participants were explicitly made aware of and tested on these payout contingencies for themselves and others before the experiment started. The experiment was conducted in a laboratory that had established a reputation of never deceiving participants, which reduces second-guessing of the experimental instructions and thereby reduces decision noise and experimenter demand effects (Hertwig & Ortmann, 2001).

In short, our studies constitute a stronger test of the hypothesis that people can detect the trustworthiness of others based on their facial appearance. In two studies, we test whether participants are more likely to entrust money to counterparts who are in fact trustworthy (Study 1, n = 131). We also compare participants’ earnings to those expected by simple decision strategies that ignore facial appearance altogether (i.e., trust at random, always trust, never trust). This allows us to test whether knowing the facial appearance of counterparts gives participants a strategic advantage in social dilemmas. In Study 2 (n = 266), we examine accuracy using an alternative experimental design. We employ an incentivized prediction task and test whether participants can accurately predict the trustworthiness of counterparts based on facial photographs.

All data and analysis scripts are available at the Open Science Framework (https://osf.io/8wejn/).¹ We report how our sample sizes were determined and all data exclusions and measures for each study.

2. Study 1

Study 1 consisted of two phases. In the first phase (n = 31), we obtained facial photographs and behavioral data from participants who acted as trustees in the trust game. In the second phase, a separate sample of participants (n = 131) made trust game decisions in the role of trustees while being matched with (and seeing photos of) the trustees of the first phase. All decisions were incentivized and both trustees and trustees received additional payments to control for social preferences (Engemann et al., 2019). We first examined whether participants relied on their counterparts’ facial appearance when making trust decisions. We tested whether participants exhibit more trust towards counterparts that are perceived as more trustworthy in two ways, by identifying the effects of variations in the perceived trustworthiness of faces, and by identifying the causal effects of trustworthiness via manipulating counterpart’s perceived trustworthiness using face morphing. The main goal of this study was to examine whether participants could accurately detect the trustworthiness of counterparts. We therefore tested (a) whether participants were more likely to transfer money to trustworthy counterparts and (b) whether knowledge of their counterparts’ facial appearance allowed them to accumulate higher earnings in the trust game than simple decision strategies that ignore facial appearance (i.e., trust at random, always trust, always distrust).

¹ Photo stimuli are not shared publicly to protect participants’ anonymity.
2.1. Methods

The study procedures were approved by the University of Zurich Ethics Committee and participants gave informed consent.

2.1.1. Stimuli (trustees)

We first collected facial photographs and behavioral strategies for a sample of trustees. Participants (n = 84) were recruited from the University of Zurich participant pool and received a fixed payment of 20 CHF (ca. $22) and additional payment that depended on their behavior in the study. At the end of the study, one round of the trust game was selected at random and participants received their earnings from that round. All decisions were therefore fully incentivized, which is an important aspect for the main studies as they reflect the true preferences of the trustees.

Participant received a written description of the “decision situation” (i.e., the trust game) and were informed that they would play five rounds with different counterparts in the role of the trustor or the trustee. In each round, both participants received an endowment of 12 CHF and the trustee could decide whether to send 10 CHF to the trustee. If the money was sent, it was tripled and transferred to the trustee. The trustee could then decide how much to send back to the trustor (between 0 and 30 CHF). We recorded trustees’ behavior with the strategy method. Trustees indicated how much they want to send back in case the trustee decided to send 10 CHF. That is, they indicated their decision without knowing whether the trustor had in fact sent anything. Participants played five rounds with anonymous counterparts and they did not receive feedback on their counterparts’ behavior, except when they found out about their earnings after the payout relevant trial was selected at the end of the experiment. This approach precludes learning and history effects from influencing decisions. The average amount of money that trustees returned to trustors (across the five rounds) constituted our measure of trustworthiness.

After completing the trust games, participants filled out a series of unrelated questionnaires and we took photographs of their faces. All photographs were taken from the same distance against a uniform background and participants were instructed to display a neutral facial expression. Similar to previous studies (Bonnefon et al., 2013), we cropped the photographs to remove all non-facial features, such as hairstyle and earrings (see Fig. 1 for an example). Sixty-three participants consented to having their photographs and behavioral data used in the study. At the end of the study, one round of the trust game was selected at random and participants, both the trustor and the trustee, received their earnings from that round.

2.1.2. Participants (trustors)

We recruited a separate sample of 273 participants from the University of Zurich participant pool. In the current study, we focus on 131 participants (M_{age} = 22.85, SD_{age} = 4.45; 45.80% female, 54.20% male) who were assigned the role of the trustor in the trust game. Participants received a fixed payment of 10 CHF (ca. $11) and were informed that they would receive an additional payment that depended on their behavior in the study. At the end of the study, one round of the trust game was selected at random and participants, both the trustor and the trustee, received their earnings from that round.

2.1.3. Procedure

Participants received the same instructions explaining the trust game as in the first phase of the study. They were informed that they would play 31 rounds in the role of the trustor with different counterparts. In each round, participants saw a photo of the trustee and decided whether to transfer nothing or 10 CHF of their 12 CHF endowment (see Fig. S1 in the Supplemental Materials). Participants also indicated what they expected the trustee to do (i.e., how much the trustee would send back in case they transferred the money) by designating amounts between 0 and 30 CHF. They indicated their confidence in the estimate on an eleven-point Likert scale ranging from “not at all certain” to “very certain.” Participants did not receive feedback on their counterparts’ behavior.

After completing the 31 rounds of the trust game, participants saw the photographs of the trustees again and rated them on various characteristics, including trustworthiness, on a seven-point scale (see Table S1 in the Supplemental Materials for a description of all measures).

2.1.4. Treatment groups

Participants were randomly assigned to one of two conditions. In the “unmodified” condition (n = 56), participants saw the original facial photographs of the trustees.

In the “modified” condition (n = 75), participants saw photographs of the same 31 trustees, but we used face morphing software to manipulate the perceived trustworthiness of trustees. Specifically, we used computer-generated face prototypes that reflect the typical appearance of a trustworthy-looking or untrustworthy-looking faces (see Fig. 1; Oosterhof & Todorov, 2008). For each trustee, we created a trustworthy-looking and an untrustworthy-looking version by morphing their face, using the software Psychomorph (Tiddeman, Burt, & Perrett, 2001), with a trustworthy-looking or untrustworthy-looking face prototype. We transformed each trustee’s face shape towards the face shape of the computer-generated prototype by 30%. This procedure created subtle differences in facial appearance (without compromising the realistic nature of the face stimuli), which affects the perceived trustworthiness of trustees (see Fig. 1). On approximately half of the 31 rounds, participants in the modified condition saw the untrustworthily-looking (vs. trustworthy-looking) version of the trustee. They only played once with each trustee, that is, they only saw one face version for each trustee.

2.1.5. Analysis strategy

Analyses were based on 1736 observations in the unmodified condition (56 participants interacting with 31 trustees) and 2325 observations in the modified condition (75 participants interacting with 31 trustees), which were analyzed separately. All analyses were conducted in R (R Core Team, 2021). We used the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) and the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2016) to estimate multilevel regression models with random intercepts and slopes. All continuous predictors were z-standardized prior to analysis (full model results are reported in the Supplemental Materials).

We followed the approach proposed by Wagenmakers (2007) to compute Bayes factors. Specifically, we estimated regression models with and without the variable of interest and computed the Bayesian information criterion (BIC), an indicator of model fit, for both models. By comparing BICs of both models, we can estimate the extent to which the variable of interest increases model fit. We converted this measure to an approximation of the Bayes factor using the following formula:

\[ BF_{10} \approx \exp \left( \frac{\text{BIC}(\text{H}_1) - \text{BIC}(\text{H}_0)}{2} \right), \]

where \( BF_{10} \) represents the Bayes factor in favor of the alternative hypothesis and \( \text{BIC}(\text{H}_1) \) and \( \text{BIC}(\text{H}_0) \) denote the fit of the models with and without the variable of interests (Wagenmakers, 2007). We used the BayesFactor package with default priors (i.e., a Cauchy distribution with a width of \( r = \sqrt{2} \); Morey and Rouder, 2018) to calculate Bayes factors for t-tests. We always display Bayes factors so that they reflect support for the favored hypothesis (i.e., \( BF_{10} \) when evidence favors the alternative hypothesis and \( BF_{01} \) when evidence favors the null hypothesis). To aid the interpretation of Bayes factors, we report the results of models with maximal random effects structure throughout the paper.

---

2 Some models only converged when we implemented simpler random effects structures. Models with maximal and simplified random effects structure yielded very similar effect size estimates and significance levels. We therefore report the results of models with maximal random effects structure throughout the paper.
factors, we classify the evidence as anecdotal, moderate, strong, very strong, or decisive (see Jeffreys, 1961).

2.1.6. Sensitivity analysis

We conducted sensitivity analyses for our main effect of interest (the relationship between participants’ trust decisions and trustees’ actual trustworthiness). We used the simr package (Green & MacLeod, 2016) in R (R Core Team, 2021) to determine the smallest effect size we were able to detect with 80% power (and α = 5%). The package provides power estimates for fixed effects in multilevel regression models. We varied the effect of interest in our model and calculated power at each level. This showed that we had 80% power to detect an odds ratio of 1.29. For a one standard deviation increase in trustworthiness, we could detect a change in the probability of trust from, for example, 50.00% to 54.29%. Thus, our design had sufficient power to detect even low levels of accuracy.

2.2. Results

2.2.1. Descriptive statistics

Across the five rounds, trustees returned an average of 6.78 CHF (SD = 6.87 CHF) of the transferred money. Nine trustees never returned anything, one trustee always returned half of the transferred money, and no trustee always returned everything. Thus, in the current sample, trust did not pay off on average, as trustees would have to return at least 10 CHF for trustees to break even. In the unmodified condition, trustees sent their 10 CHF endowment 54.09% of the time. Eleven participants (19.64%) never trusted whereas nine participants (16.07%) always trusted. In the modified condition, trustees sent their 10 CHF endowment 52.77% of the time. Fourteen participants (18.67%) never trusted whereas thirteen participants (17.33%) always trusted.

2.2.2. Manipulation check

To test whether our morphing manipulation affected the perceived trustworthiness of trustees in the modified condition, we compared participants’ trustworthiness ratings of the two face versions. The morphed trustworthy faces (M = 3.25, SD = 0.40) were rated as significantly more trustworthy (with decisive evidence in favor of the alternative hypothesis) than the original trustworthy faces (M = 2.32, SD = 0.42), t(30) = 16.48, p < .001, d = 2.96, BF10 = 3.92 × 1013. Moreover, the morphed trustworthy faces were rated as significantly more trustworthy (with decisive evidence in favor of the alternative hypothesis) than the original faces (M = 2.89, SD = 0.44), t(30) = 7.96, p < .001, d = 1.43, BF10 = 1.84 × 106, while the morphed untrustworthy faces were rated as significantly less trustworthy (with decisive evidence in favor of the alternative hypothesis) than the original faces, t(30) = 10.92, p < .001, d = 1.96, BF10 = 1.53 × 1015. Thus, our morphing procedure successfully manipulated the perceived trustworthiness of trustees.

2.2.3. Reliance on facial appearance

First, we examined whether participants who saw the unmodified photographs relied on the facial appearance of trustees when deciding whom to trust. We estimated a multilevel regression model with random intercepts and slopes per participant in which we regressed participants’ trust behavior (0 = did not transfer endowment, 1 = transferred endowment) on their trustworthiness ratings. This yielded a positive effect with very strong evidence in favor of the alternative hypothesis, β = 0.864, SE = 0.173, OR = 2.37, 95% CI [1.65, 3.58], p < .001, BF10 = 41.23 (see Fig. 2A, Table S2). Participants were more likely to trust when they perceived their counterparts as trustworthy.

The positive relationship between perceived facial trustworthiness and trust behavior may also reflect a consistency effect. Rather than relying on the facial appearance of counterparts when making trust decisions, participants may have rated counterparts as more trustworthy because they trusted them. We addressed this alternative explanation in two ways. First, we computed average trustworthiness ratings of counterparts across all participants. Using this average trustworthiness rating instead of individual ratings, perceived trustworthiness was again positively related to the probability of trust (with decisive evidence in favor of the alternative hypothesis), β = 0.556, SE = 0.093, OR = 1.74, 95% CI [1.44, 2.16], p < .001, BF10 = 601.0 (see Fig. 2B, Table S3).

Second, to estimate the causal effect of facial appearance on trust decisions more directly, we analyzed the effect of our morphing manipulation on participants’ behavior in the modified condition. Increasing (vs. decreasing) the facial trustworthiness of trustees had a positive influence on the probability of trust (with decisive evidence in favor of the alternative hypothesis), β = 0.994, SE = 0.149, OR = 2.70, 95% CI [1.98, 3.69], p < .001, BF10 = 2314 (see Fig. 2C, Table S4). Together, these results show that participants relied on the facial appearance of counterparts when making trust decisions.

2.2.4. Trustworthiness detection

The above results show that participants relied on facial trustworthiness judgments when making trust decisions. But is this wise given that decisions had real financial consequences? We therefore asked whether participants were able to detect the true trustworthiness of
counterparts based on the facial photographs. To address our main research question, we tested whether participants who saw the unmodified photos were more likely to trust counterparts that were actually more trustworthy. We regressed trust behavior on the behavioral (not the perceived) trustworthiness of counterparts (i.e., the average amount of money that trustees had returned to trustors), which did not yield a significant effect and very strong evidence in favor of the null hypothesis, $\beta = 0.048, SE = 0.075, OR = 1.05, 95\% CI [0.89, 1.23], p = .52, BF_{10} = 34.50$ (see Table S4, Model 1). Thus, results showed that participants were not able to detect the true trustworthiness of counterparts based on trustworthiness inferences from photographs.

This last result suggests that reliance on the facial appearance of counterparts did not pay off, which we tested directly by comparing our participants' average performance to that of other decision strategies. If knowledge about the facial appearance of trustees actually gives trustors a strategic advantage, then participants' earnings across the 31 rounds should be higher than the earnings of a person who trusts at random. Participants' earned an average of 257.1 CHF across the 31 rounds ($SD = 40.69$ CHF). Crucially, participants' earnings were not significantly higher (with substantial evidence in favor of the null hypothesis) than the earnings of a trustor choosing at random ($M = 260.1$ CHF), $t(55) = 0.55, p = .58, d = 0.07, BF_{10} = 5.93$ (see Fig. 3).

Another simple but potentially viable strategy for making trust decisions would be to (a) estimate whether trust will pay off on average (in the current context, whether trustees will on average return more than 10 CHF) and (b) always trust if it does or always distrust if it does not. Participants' earnings were higher than those of an always-trust strategy ($M = 210.0$ CHF) with decisive evidence in favor of the alternative hypothesis, $t(55) = 8.66, p < .001, d = 1.16, BF_{10} = 1.22 \times 10^9$, but lower than those of an always-distrust strategy ($M = 310.0$ CHF) with decisive evidence in favor of the alternative hypothesis, $t(55) = 9.73, p < .001, d = 1.30, BF_{10} = 5.25 \times 10^{10}$ (see Fig. 3). Together, these results suggest that having access to the facial appearance of trustees did not give participants a strategic advantage. In fact, knowledge about the base rate of trustworthiness in the current sample of trustees (i.e., the fact that trust did not pay off on average) and a resulting strategy of consistent distrust would have resulted in higher earnings.

### 2.2.5. Moderators of accuracy

Next, we tested the hypothesis that accuracy in trustworthiness detection only emerges under specific conditions. Contrary to Tognetti et al. (2013), who found evidence for increased accuracy when participants were judging male but not female trustees, accuracy did not vary as a function of trustees' gender (with strong evidence in favor of the null hypothesis), $\beta = -0.036, SE = 0.163, OR = 0.96, 95\% CI [0.69, 1.34], p = .83, BF_{10} = 40.74$ (Table S4, Model 2). We also explored whether accuracy varied as a function of trustees' gender or trustees' confidence in the accuracy of their expectations of reciprocity (Table S4, Models 3 and 4), but found no significant results and very strong to decisive evidence in favor of the null hypothesis.

### 2.2.6. Additional analyses

Two additional variables were recorded that provide additional insights about participants' knowledge of the trustworthiness of the trustees: participants' expectancy of reciprocity and their explicit trustworthiness ratings. We first analyzed whether participants' reciprocity expectation was associated with their counterparts' actual trustworthiness. The relationship between how much participants expected trustees to return and how much they actually returned was not significant with decisive evidence in favor of the null hypothesis, $\beta = 0.050, SE = 0.104, 95\% CI [-0.151, 0.249], p = .63, BF_{10} = 142.1$ (Table S5). There was a significant positive association between explicit trustworthiness ratings and trustees' actual trustworthiness, $\beta = 0.078, SE = 0.027, 95\% CI [0.028, 0.136], p = .004, BF_{10} = 9.91$ (Table S6). It should be noted that this relationship was very small and Bayesian analyses indicated...
substantial evidence in favor of the null hypothesis.

3. Study 2

Results of Study 1 suggest that participants were not able to detect the trustworthiness of counterparts based on facial photographs. However, decisions in the trust game may be motivated by considerations other than the expected trustworthiness of counterparts. For instance, people may transfer money not because they think that their counterpart will reciprocate trust, allowing them to maximize their earnings, but because there is an injunctive norm to trust and to not question a counterpart’s character (Dunning, Anderson, Schlösser, Ehlebracht, & Fetchenhauer, 2014). In Study 2, we therefore examined trustworthiness detection accuracy with an incentivized prediction task, in which participants’ earnings were tied to the accuracy of their predictions. Participants viewed the same cropped images used in the unmodified condition of Study 1 and predicted the trustworthiness of trustees. We also examined detection accuracy for uncropped images in a separate condition.

3.1. Methods

The study procedures were approved by the University of Zurich Ethics Committee and participants gave informed consent.

3.1.1. Participants

We recruited a sample of 266 participants from the University of Zurich participant pool (Mage = 22.31, SDage = 3.38; 49.25% female, 51.75% male). Participants received a fixed payment of 20 CHF (ca. $22) and a variable payment that depended on the accuracy of their guesses (see the Supplemental Materials for the exact payoff formula).

3.1.2. Procedure and treatment groups

Participants received written instructions that explained the trust game played by the stimulus group. They were asked to view photographs of these players and to guess the behavior of the players as accurately as possible. The instructions were followed by a comprehension test that tested whether participants had understood the game and the manner in which their own payments related to their guessing accuracy. Participants could not begin the study until all comprehension questions had been answered correctly. When viewing images of trustees, participants were asked to guess in what percentage of rounds the person sent 10 CHF, on a scale that ranged from 0% to 100%. When viewing images of trustees, participants were asked to guess the average amount that the trustee sent back, on a scale that ranged from 0 CHF to 30 CHF. For each guess, participants also indicated their confidence in the estimate on an 11-point Likert scale ranging from 0 CHF to 100 CHF. Participants were asked to explicitly predict the trustworthiness of trustees who viewed the uncropped images and found similar results, $\beta = -0.104, SE = 0.288, 95\% CI [-0.654, 0.489], p = .72, BF_{01} = 40.74$ (Table S8, Model 1). Just like in the context of the trust game, participants that were incentivized to explicitly predict the trustworthiness of trustees were not able to do so.

3.1.3. Analysis strategy

We followed the same analysis strategy as in Study 1. For all tests, we report the results of frequentist and Bayesian analyses. We estimated cross-classified multilevel regression models with random intercepts and slopes per participant and trustee (full model results are reported in the Supplemental Materials).

3.1.4. Sensitivity analysis

We again conducted sensitivity analyses for our main effect of interest (the relationship between predicted and actual trustworthiness in the cropped and uncropped conditions). For participants in the cropped condition, we had 80% power to detect an effect of 0.10. In other words, for a one-point increase in actual trustworthiness, we could detect a 0.10-point increase in predicted trustworthiness. For participants in the uncropped condition, we had 80% power to detect an effect of 0.14. In other words, for a one-point increase in actual trustworthiness, we could detect a 0.14-point increase in predicted trustworthiness. Thus, our design had sufficient power to detect even low levels of accuracy.

3.2. Results

3.2.1. Trustworthiness detection

Were participants able to detect the trustworthiness of counterparts? We estimated a multilevel regression model with random intercepts and slopes per participant, in which we regressed the predicted reciprocation rate of trustees on their actual reciprocation rate. For participants who viewed the same cropped images as participants in Study 1, there was no significant relationship between how much participants predicted trustees to return and how much they had actually returned (with very strong evidence in favor of the null hypothesis), $\beta = -0.043, SE = 0.263, 95\% CI [-0.456, 0.559], p = .87, BF_{01} = 64.20$ (Table S7, Model 1). We estimated a separate model for participants who viewed the uncropped images and found similar results, $\beta = -0.104, SE = 0.288, 95\% CI [-0.654, 0.489], p = .72, BF_{01} = 40.74$ (Table S8, Model 1). Just like in the context of the trust game, participants that were incentivized to explicitly predict the trustworthiness of trustees were not able to do so.

3.2.2. Moderators of accuracy

We again tested the hypothesis that accuracy in trustworthiness detection only emerges under some conditions. Accuracy did not vary as function of trustees’ gender (with substantial to very strong evidence in favor of the null hypothesis) in the cropped condition, $\beta = -0.283, SE = 0.496, OR = 0.75, 95\% CI [0.27, 2.10], p = .57, BF_{01} = 30.47$ (Table S7, Model 2), and in the uncropped condition, $\beta = -0.791, SE = 0.595, OR = 0.45, 95\% CI [0.14, 1.49], p = .20, BF_{01} = 9.01$ (Table S8, Model 2). We also explored whether accuracy varied as a function of participants’ gender (Model 3, Tables S7 and S8) or participants’ confidence in the accuracy of their expectations of reciprocity (Model 4, Tables S7 and S8), but found no significant results and very strong to decisive evidence in favor of the null hypothesis.

3.2.3. Additional analyses

To replicate results from Study 1 outside the context of a trust game, we again analyzed the accuracy of explicit trustworthiness ratings of the facial photographs. We did not find a significant association between trustworthiness ratings and the actual reciprocation rate of trustees (with decisive evidence in favor of the null hypothesis) for participants who viewed the cropped images, $\beta = 0.066, SE = 0.102, 95\% CI [-0.155, 0.294], p = .53, BF_{01} = 136.6$ (Table S9), and for participants who viewed the uncropped images, $\beta = 0.006, SE = 0.114, 95\% CI [-0.237, 0.237], p = .96, BF_{01} = 109.0$ (Table S10). Together, these results suggest that participants were not able to predict the trustworthiness of counterparts based on facial photographs.

4. General discussion

People spontaneously rely on the facial appearance of strangers when deciding whether they can be trusted to cooperate in social interactions (Todorov, Olivola, et al., 2015). But can people actually detect the trustworthiness of strangers based on their facial appearance? Prior studies have yielded mixed results and the question remains the subject of vigorous debate (Bonnefon et al., 2017; Todorov, Funk, & Olivola, 2015). Yet, the empirical evidence on the topic is limited. Many
studies were based on the same set of stimuli, which limits the generalizability of findings (Bonnefon et al., 2013; De Neys et al., 2015, 2017). Conversely, studies providing evidence against accuracy relied on statistical techniques that cannot quantify evidence in favor of such a null hypothesis, which complicates the interpretation of results (Effer, 2013). Diagnostic cues are available (Jaeger et al., 2019) and when decisions are highly consequential (Wilson & Rule, 2015). Studies were based on the perceived trustworthiness of counterparts when making trust decisions. However, on average, participants failed to entrust money to counterparts that were actually more trustworthy. Bayesian analyses yielded very strong support for the null hypothesis indicating that our participants were not able to accurately detect the trustworthiness of their interaction partners. We also found that participants’ earnings were not higher than the expected earnings of a decision strategy that trusts at random. This suggests that knowledge of their counterparts’ facial appearance did not give participants a strategic advantage. In fact, participants would have earned more by consistently distrusting all counterparts, as trust did not pay off in the current sample.

Previous studies found evidence in favor of detection accuracy only under specific conditions, and these conditions varied across studies (Bonnefon et al., 2013; Tognetti et al., 2013; Verplaetse et al., 2007). Here, we tested these proposed moderators, but found no evidence for better-than-chance trustworthiness detection (a) for male or female counterparts, (b) when making trust decisions or when providing explicit trustworthiness ratings, and (c) when viewing cropped images (in which all non-facial features were removed) or uncropped images. In sum, our results provide consistent evidence against accuracy in trustworthiness detection from faces across various conditions.

Previous investigations have shown that trustworthiness impressions guide decision-making in many domains, including legal sentencing, personnel selection, and financial decision-making (Olivola et al., 2014). People even rely on trustworthiness impressions from faces when more diagnostic cues are available (Jaeger et al., 2019) and when decisions are highly consequential (Wilson & Rule, 2015). Future studies should explore whether some people are more prone to the biasing influence of first impressions and, importantly how biases could be mitigated (for first attempts, see Chua & Freeman, 2021; Jaeger, Todorov, Evans, & van Beest, 2020; Shen & Ferguson, 2021). An important future task in this line of research will be to delineate how difficult it is to override these biases, particularly when other more reliable information sources are available that may require more cognitive effort to process.

4.1. Limitations and future directions

Several limitations and constraints on the generalizability of the current results should be mentioned. Our results were based on samples of relatively young decision-makers from the University of Zurich. Additional studies are needed to examine the generalizability of our findings with larger and more diverse samples of both targets and raters. Further studies should also examine the accuracy of trustworthiness impressions using varying types of stimuli. Cropped images, in which all non-facial aspects are removed, ensure that impressions are based on the facial features of counterparts. However, they represent only a relatively specific facet of the kinds of stimuli that people encounter in everyday interpersonal interactions. Accuracy may be better than chance when people have access to additional cues. For instance, previous findings suggest that people may be able to identify cooperative interaction partners with greater-than-chance accuracy after brief interactions (Brosig, 2002; DeSteno et al., 2012; Frank, Gilovich, & Regan, 1993; Reed, Zeglen, & Schmidt, 2012; but see Manson, Gervais, & Kline, 2013; McCullough & Reed, 2016). Ultimately, we believe that studies using a wide range of different stimuli are needed to map the accuracy of trustworthiness decisions under varying conditions.

We investigate one such condition here, namely the accuracy of trustworthiness judgments when judgments are solely based on facial features. This approach is informative for two reasons. First, even though people often have access to other cues, which may allow them to make more accurate judgments, there are also many situations in which a person’s facial appearance is either one of the only cues or a particularly salient cue. People often engage with strangers and, in the first moments of the interaction, tend to judge them solely based on their appearance. Moreover, facial photographs are a common feature of many decision-making environments, including social media platforms (e.g., Twitter), professional networking sites (e.g., LinkedIn), and the sharing economy (e.g., Airbnb). Second, ample evidence suggests that people rely on facial appearance, even when they have access to other cues (Jaeger et al., 2019; Olivola et al., 2014). To determine whether reliance on facial appearance helps or hinders people in making accurate predictions, the accuracy of judgments that are solely based on facial appearance needs to be isolated. This requires a highly controlled and standardized study design, such as the one used in the current experiments, to ensure that judgments are based on the cue in question (Cox et al., 2015).

Declarations of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.evolhumbehav.2022.04.004.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1). https://doi.org/10.18637/jss.v067.i01
Berg, J., Dickhaut, K., & McCabe, K. (1995). Trust, reciprocity, and social history. Games and Economic Behavior, 10(1), 122–142.
Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. Journal of Experimental Psychology: General, 142(1), 143–150. https://doi.org/10.1037/a0028930
Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2017). Can we detect cooperators by looking at their face? Current Directions in Psychological Science, 26(3), 276–281. https://doi.org/10.1177/0963721416693352
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
Brosig, J. (2002). Identifying cooperative behavior: Some experimental results in a prisoner’s dilemma game. Journal of Economic Behavior & Organisation, 47(3), 275–290. https://doi.org/10.1016/S0167-2681(01)00211-6
Camerer, C., & Mobbs, D. (2017). Differences in behavior and brain activity during hypothetical and real choices. Trends in Cognitive Sciences, 21(1), 46–56. https://doi.org/10.1016/j.tics.2016.11.001 Chua, K. W., & Freeman, J. B. (2021). Facial stereotype bias is mitigated by training. Social Psychological and Personality Science, 12(7), 1335–1344. https://doi.org/10.1177/1948550620972550
Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In The adapted mind (pp. 163–228). https://doi.org/10.1098/rstb.2006.1997
Cox, W. T. L., Devine, P. G., Bischmann, A. A., Hyde, J. S., Cox, W. T. L., Devine, P. G., … Hyde, J. S. (2015). Inferences about sexual orientation: The roles of stereotypes, faces, and the Gaydar myth. The Journal of Sex Research, 53(2), 157–171. https://doi.org/10.1080/00224499.2015.1017714
Dawes, R. M. (1980). Social dilemmas. Annual Review of Psychology, 31, 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125
De Neys, W., Hopfensitz, A., & Bonnefon, J. F. (2017). Split-second trustworthiness detection from faces in an economic game. Experimental Psychology, 64, 231–239. https://doi.org/10.1016/1813-3166(a000367)
Evolution and Human Behavior xxx (xxxx) xxx

B. Jaeger et al.

Manson, J. H., Gervais, M. M., & Kline, M. A. (2013). Defectors cannot be detected during mixed effects models (R package version 2.0-32).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). Mixed effects models for continuous outcomes with R: The lmerTest package.

Dunning, D., Anderson, J. E., & Schmidt, J. (2013). The role of context in the evolution of one-shot cooperation: An experiment. Ethology and Sociobiology, 24, 273–284.

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). The evolution of one-shot cooperation: A residential perspective. Behavioral and Brain Sciences, 16, 639–689.

Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceivers and target characteristics in person perception. Journal of Personality and Social Psychology, 113(4), 513–529. https://doi.org/10.1037/pspa0000090

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: a methodological challenge for psychologists? Economic Psychology, 47, 17–22. https://doi.org/10.1016/j.econpsych.2015.01.002

DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., & Lee, J. J. (2019). Explaining the persistent influence of facial cues in social decision-making. Proceedings of the National Academy of Sciences, 116(8), 3427–3432. https://doi.org/10.1073/pnas.1814779116

Efferson, C., & Vogt, S. (2013). Viewing men’s faces does not lead to accurate predictions of trustworthiness. Evolution and Human Behavior, 34(2), 126–131. https://doi.org/10.1016/j.evolhumbehav.2013.08.002

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. Trends in Cognitive Sciences, 18(11), 566–570. https://doi.org/10.1016/j.tics.2014.09.007

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. Proceedings of the National Academy of Sciences, 105(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Efferson, C., & Vogt, S. (2013). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. Journal of Personality and Social Psychology, 104(3), 409–426. https://doi.org/10.1037/a0031050

Shen, X., & Ferguson, M. J. (2021). How resistant are implicit impressions of facial trustworthiness? When new evidence leads to durable updating. Journal of Experimental Social Psychology, 97, Article 104219. https://doi.org/10.1016/j.jesp.2021.104219

Tiddeman, B., Burt, M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. Computer Graphics and Applications, IEEE, 2(5), 32–50.

Todorov, A., Funk, F., & Olivola, C. Y. (2015). Response to Bonnefon et al.: Limited ‘kernels of truth’ in facial inferences. Trends in Cognitive Sciences, 19(8), 422–424. https://doi.org/10.1016/j.tics.2015.05.002

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. Annual Review of Psychology, 66(1), 519–545. https://doi.org/10.1146/annurev-psycho-113011-143831

Tognetti, A., Berticat, C., Raymond, M., & Faurie, C. (2013). Is cooperativeness readable of faces? A cross-cultural approach. Evolution and Human Behavior, 34(6), 427–432. https://doi.org/10.1016/j.evolhumbehav.2013.08.002

Verplanken, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: The sequel: A kernel of truth in predictive cheating detection. Evolution and Human Behavior, 28(4), 260–271. https://doi.org/10.1016/j.evolhumbehav.2007.04.006

Vangenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. Psychonomic Bulletin & Review, 14(5), 779–804. https://doi.org/10.3758/PBR.14.5.779

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

Yamagishi, T., Tanida, S., Mashima, R., Shimoma, E., & Kanazawa, S. (2003). You can judge a book by its cover: Evidence that cheaters may look different from cooperators. Evolution and Human Behavior, 24(4), 290–301. https://doi.org/10.1016/S1090-5138(03)00035-7

Zylberzteje, A., Babutsdiez, Z., & Hanaki, N. (2020). Preferences for observable information in a strategic setting: An experiment. Journal of Economic Behavior and Organization, 170, 268–285. https://doi.org/10.1016/j.jebo.2019.12.009