SPLIT-SECOND TRUSTWORTHINESS DETECTION FROM FACES IN AN ECONOMIC GAME

Wim De Neys[1] , Astrid Hopfensitz[2] , & Jean-François Bonnefon[3]

[1] LaPsyDE (CNRS Unit 8240), Sorbonne - Paris Descartes University, Paris, France

[2] Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France

[3] Toulouse School of Economics, Center for Research in Management, University of Toulouse, Toulouse, France

Word count: 4,998

Mailing address: Wim De Neys
LaPsyDÉ (Unité CNRS 8240, Université Paris Descartes)
Sorbonne - Labo A. Binet
46, rue Saint Jacques
75005 Paris
France

wim.de-neys@parisdescartes.fr

ABSTRACT

Economic interactions often imply to gauge the trustworthiness of others. Recent studies showed that when making trust decisions in economic games, people have some accuracy in detecting trustworthiness from the facial features of unknown partners. Here we provide evidence that this face-based trustworthiness detection is a fast and intuitive process by testing its performance at split-second levels of exposure. Participants played a Trust game, in which they made decisions whether to trust another player based on their picture. In two studies, we manipulated the exposure time of the picture. We observed that trustworthiness detection remained better than chance for exposure times as short as 100 ms, although it disappeared with an exposure time of 33ms. We discuss implications for ongoing debates on the use of facial inferences for social and economic decisions.

INTRODUCTION

The success of the human species is rooted in our remarkable ability to cooperate, which implies to accept social risk on the basis of trust (Rand & Nowak, 2013; Tomasello, 2009). Indeed, not everyone trusted will later prove trustworthy, as the victims of Bernard Madoff and other scammers will testify. Because trusting others puts one at the risk of being exploited by cheaters, it is important to accurately assess the trustworthiness of others. Given the importance of trust for human interactions, research on trusting and trustworthy behavior spans a large range of disciplines, including psychology, economics, and sociology (e.g., Chang, Doll, van 't Wout, Frank, & Sanfey, 2010; Delhey & Newton, 2003; Evans & Krueger, 2009). As a consequence, there is large variety of paradigms for measuring trust, trustworthiness, and trustworthiness detection. Among the most common paradigms is the Trust Game (Berg, Dickhaut, & McCabe, 1995), which allows researchers to capture trusting decisions in a controlled but reasonably realistic economic setting.

The general structure of this simple game involves two players, A and B. Player A (the *Investor*) is endowed with an initial amount of money ($10), and can choose whether to keep that money or transfer it to player B (the *Trustee*). If player A decides to transfer the money, the experimenter multiplies it by a factor larger than 1 (this factor is heretofore assumed to be 3) before giving it to Player B. Player B then decides how much of the resulting $30, if any, should be returned to Player A. One extreme decision would be to return the whole $30 to Player A, in which case Player B would win nothing. Another extreme decision would be to return nothing to Player A, in which case Player B would pocket the whole $30. In addition any split of the total amount between Player A and Player B is possible. Both players are completely informed about the procedure but cannot communicate during the game. In essence, the Investor needs to decide whether or not to trust the Trustee to return some money to him, and the Trustee can decide whether to honor or to abuse this trust. Trustworthiness detection can be measured in this game by comparing the decisions of the Investor to the strategies of the Trustees: an Investor would demonstrate perfect trustworthiness detection by transferring money

to all Trustees who honor trust, and not transferring any money to Trustees whose strategy is to abuse trust.

Available evidence suggests that Investors strongly rely on the facial appearance of the Trustees to make their decisions (e.g., Engell, Haxby, & Todorov, 2007; van 't Wout & Sanfey, 2008; Rezlescu, Duchaine, Olivola, & Chater, 2012). In particular, Investors are more likely to transfer money to Trustees whose faces are rated as trustworthy by an independent group of judges—the correlation between trustworthiness ratings and transfer decisions being typically in the vicinity of .80 (Van t' Wout and Sanfey, 2008). Additionally, numerous studies showed that facial trustworthiness ratings had a high inter-rater reliability, suggesting that everybody mostly agree about who looks trustworthy and who does not (e.g., Freeman, Stolier, Ingbretsen, & Hehman, 2014; Todorov, Pakrashi, & Oosterhof, 2009). These results do not speak to the issue of trustworthiness *detection,* though. We know that Investors place a strong emphasis on Trustees' faces, but does it allow them to discriminate trustworthy Trustees from untrustworthy Trustees?

The available evidence would seem to suggest that while explicit trustworthiness ratings of faces are inaccurate, the decisions that Investors make in a trust game show some minimal but observable level of accuracy (Olivola et al., 2014; Bonnefon, Hopfensitz, & De Neys, 2015, 2017). On the one hand, although people mostly agree on who looks trustworthy and who does not, their evaluations are far from accurate. Individuals who are judged to look trustworthy do not necessarily behave in a trustworthy manner (Efferson & Vogt, 2013; Rule, Krendl, Ivcevic, & Ambady, 2013; Todorov, Olivola, Dotsch, Mende-Siedlecki, 2014; Olivola, Funk, & Todorov, 2014; Wilson & Rule, 2015; but see Little, Jones, De Bruine, & Dunbar, 2013; Tognetti, Berticat, Raymond, & Faurie, 2013). For example, Rule et al. (2013) asked participants to rate the trustworthiness of pictures of business executives who had been convicted of financial crimes and others who had not been convicted. Trust ratings for convicted and non-convicted executives did not differ. Likewise, pictures of people who cheated on a test were not rated differently than the pictures of individuals who did not cheat. These and many other examples indicate that explicit ratings of trustworthiness are invalid (Olivola et al., 2014).

On the other hand, a number of studies showed that actual decisions in the trust game, in contrast to mere trustworthiness ratings, showed a small but robust level of accuracy (e.g., Bonnefon, Hopfensitz, & De Neys, 2013; De Neys, Hopfensitz, & Bonnefon, 2013, 2015; Stirrat & Perret, 2010). In these studies, Investors typically played trust games (for real payoffs) with a series of Trustees, whose strategies were recorded in advance, and whose faces were shown to the Investors. The critical observation was that Investors were slightly more likely to transfer money to Trustees who honored trust than to Trustees who abused trust. This indicates that people have some minimal accuracy in reading trustworthiness from the faces of unknown adults when they have to make a decision in the trust game. Importantly, Bonnefon et al. (2013, Study 4) collected explicit ratings of facial trustworthiness using the same pictures that they used in the trust game. They found that ratings of facial trustworthiness did not discriminate between Trustees that honored trust and Trustees that did not, even though the latter received significantly less transfers in the Trust Game. Furthermore, they observed that trustworthiness ratings only explained 18% of the variance in the transfer decisions that discriminated between Trustees who did and did not honor trust.

Why do face-based decisions in the trust game show some small level of accuracy, whereas ratings of facial trustworthiness do not? We suggest to approach this issue by investigating the time course of face-based trust decisions, in the context of existing research on the time course of trustworthiness ratings. Indeed, ratings of facial trustworthiness appear to be extremely fast: ratings provided after a 100 ms exposure to a face strongly correlate with ratings provided after unlimited exposure time (Willis & Todorov, 2006). Such results led to the hypothesis that impressions of facial trustworthiness are automatic and intuitive in nature, and do not require deliberate reflection (Olivola & Todorov, 2010; Willis & Todorov, 2006; Todorov, Mandisorlza, Goren, & Hall, 2005). This may suggest an explanation for the discrepancy in accuracy between ratings and decisions: Trustworthiness ratings would be fast, intuitive, and unreliable; whereas trusting decisions would be slower, more deliberate, and more accurate as a result. Indeed, experiments which supported the accuracy of face-based decisions used long exposure time, for example more than five seconds in Bonnefon et al. (2013) and De Neys et al.

(2014, 2015), which would allow deliberate processes to run their course (e.g., De Neys & Bonnefon, 2013; Evans & Stanovich, 2013).

This account affords a testable prediction. Given that deliberate processing requires more time than automatic processing, the accuracy of face-based trusting decisions should disappear with shorter exposure time such as 100 ms (for which trustworthiness ratings retain their inter- and intra-rater reliability). There are reasons to doubt that face-based decisions result from a deliberative thought-process, though. In particular, Bonnefon et al. (2013) observed that burdening Investors' executive resources with a concurrent task did not impair trustworthiness detection. Because executive resources are required for deliberate processing (De Neys, 2006a, Kahneman, 2011), this result suggests that trustworthiness detection in the trust game is largely automatic—and given this automatic nature, it is reasonable to assume that it can operate under very short exposure times (De Neys, 2006b; De Neys & Bonnefon, 2013; Evans & Stanovich, 2013). Given these contrasting predictions, Study 1 was conducted as a test of whether trustworthiness detection in the trust game would survive or disappear with a 100 ms exposure time to the faces of the Trustees.

STUDY 1

METHOD

Participants

A total of 41 Belgian university students (25 females, mean age = 21.3, SD = 1.8, range = 19-27) participated voluntarily. Participants were informed that they could receive a monetary compensation depending on their performance in the game (0 euro, 4 euros, or 6 euros).

Material and procedure

Trust game. Participants were tested individually in a quiet room. Upon entering the room participants were familiarized with the rules of the Trust game. Each participant played a

total of 14-rounds in the role of Investor, with 14 different Trustees. Note that while the words 'Investor' and 'Trustee' are used for clarity here, they were never actually used in the experiment. The Trustee was simply called 'the other player'. Each round had the same structure: Participants were endowed with a sum of 4 euros and had to decide whether to keep the endowment, or to transfer that endowment to a Trustee, whose picture appeared on the screen. In case the endowment was transferred, it was multiplied by three, and the Trustee had to decide whether to keep the whole 12 euros, to return 6 euros to the Investor, or to return 4 euros to the Investor. We refer to these strategies as the Abuser, Cooperator, and Neutral strategies, respectively. These terms were not mentioned to the participants. The participants were informed that each Trustee had already recorded his or her strategy. Participants were also informed that one round would be randomly selected after the study, and that they would receive whatever money they made in that round.

Trustors' payoffs were based on a pairing with a randomly selected Trustee. Trustees did not receive any payoff in the current set of studies (they did receive their payoff in the previous study in which their picture had been taken and their strategy recorded, see Centorrino, Djemai, Hopfensitz, Milinski, & Seabright, 2015). This implied the use of deception in that Trustors believed that the Trustee's pay-off was contingent on the Trustor's own behavior (see Hertwig & Ortmann, 2008, for a critical discussion of the use of deception in economic experiments). Note also that in our variant of the Trust game the Trustee received a fixed show-up fee but no further initial endowment in the game (e.g., see Glaeser, Laibson, Scheinkman, & Soutter, 2000). This can affect the absolute level of trust (e.g., due to inequality aversion Trustors may be more likely to transfer money overall, e.g., Ciriolo, 2007). However, given that our core interest lies in the accuracy of trustworthiness detection (i.e., relative difference in transfer to Cooperators vs Abusers) this design feature should not bias results.

The crucial manipulation in the current study was the presentation or exposure time of the picture of the Trustee (5500 ms or 100 ms). Each round started with the presentation (1500 ms) of the number of the round (e.g., "Round 1") that was going to be played. Next, a fixation cross (1000 ms) was presented in the middle of the screen, followed by the picture of the Trustee (5500 ms or 100 ms presentation). Participants were randomly allocated to one of the two

exposure time conditions. Pictures were presented in random order for each participant. Participants indicated whether they wanted to transfer money to the Trustee by pressing one of two keys, after which they moved on to the next round, without receiving feedback about the strategy of the Trustee.

Before the 14-round Trust game started participants could play one practice round to familiarize them with the procedure. After having played the 14 rounds, participants were asked for an estimate of the overall proportion of Trustees that they believed would return nothing and their own strategy had they played the role of Trustee. The answers to these questions were not further analyzed.

Trustee pictures. The Trustees shown to the participants had recorded their strategy and were incentivized in the context of a previous study (Centorrino et al., 2015). In this initial study, 84 young adults played the role of Trustee, and recorded a movie introducing themselves. From each of these movies, a research assistant blind to the strategies of the Trustees extracted one frame in which the Trustee had the most neutral expression. Each picture was then cropped (left and right facial boundaries, chin and top of the eyebrows) to minimize display of clothing or hairstyle, and turned to black and white (Figure 1). The trustworthiness detection study of Bonnefon et al. (2013) used a set of 60 of these pictures. To keep participants focused and motivated with the present brief exposure times we selected a subset of 14 pictures: 6 pictures of Abusers (3 males, 3 females), 6 pictures of Cooperators (3 males, 3 females), and 2 pictures of Neutral players (1 male, 1 female) that were used as fillers. Because the current study investigates a potential negative moderator (exposure time) of the original trustworthiness detection effect, we maximized power by selecting pictures of Cooperators and Abusers that were highly discriminated in the original studies with 5500 ms exposure time. That is, we selected the 3 male Cooperators and 3 female Cooperators who were best discriminated as such in previous studies, and the 3 male Abusers and 3 female Abusers who were best discriminated as such in previous studies (e.g., see De Neys et al., 2015).

RESULTS

For each participant we calculated the average transfer rate to cooperators and abusers. These averages were subjected to a 2 (exposure time, between-subjects; 100 ms or 5500 ms) x 2 (Trustee strategy, within-subjects; cooperator or abuser) mixed model ANOVA.[1] Figure 2 (top panel) shows the results. As Figure 2 indicates, there was a main effect of Trustee strategy, $F(1, 39) = 5.01$, $p < .05$, $\eta_p^2 = .11$. Overall, participants were about 10% more likely (95% confidence interval: 01-19) to transfer their endowment to Trustees who honored trust than to trustees who abused trust. Critically, neither the main effect of exposure time, $F(1, 39) < 1$, nor the interaction with Trustee strategy, $F(1, 39) < 1$, reached significance. As Figure 2 (top panel) illustrates, reducing exposure time from the 5500 ms to 100 ms had no detectable impact on the accuracy of the trust decisions (see the results of Study 2 for an overall Bayes factor analysis further supporting this conclusion).

Results were robust to the inclusion of the sex of the trustee as an additional within-subject predictor. The sex of the trustee did not impact transfer rates and did not significantly interact with any other factor. Finally, we computed a sensitivity index *d'* for participants in the 100ms and 5500ms conditions, in order to capture their ability to discriminate cooperators. The index *d'* is defined as Z(*Tc*) - Z(*Ta*), where *Tc* and *Ta* denote the proportion of trusting decisions made when playing with cooperators and abusers (respectively), and where Z returns the inverse of the standard normal cumulative distribution. Because *d'* is undefined when either *Ta* or *Tc* is 0 or 1, we followed convention and replaced these values with 1/12 and 1-1/12 (respectively), where 12 corresponds to the number of decisions made by each participant. The index *d'* took the values 0.25 and 0.27 in the 100ms and 5500ms conditions, respectively, and these values were not statistically different ($p = .94$, the 95% confidence interval for the difference in *d'* being -0.52–0.48).

Study 2 was conducted to consolidate and generalize the results obtained in Study 1. Importantly, Study 2 introduced a masking procedure ensuring that faces could only be pro-

[1] The analysis does not include transfers to neutral trustees. There are no theory-driven predictions about these transfers, and their estimation is noisy given that they are only based on two pictures. The 95%-confidence interval for transfers to neutral trustees was 07—53 in the 5500 ms condition, and 31—79 in the 100 ms condition.

cessed for the precise, intended exposure time. Because an image will be briefly stored in visual sensory memory, it can be processed slightly longer than the mere exposure time (Sperling, 1960). The masking procedure consists of presenting a second, masking image immediately after the Trustee picture, which disrupts further visual processing of the picture. As such, it guarantees that processing time will not be unintentionally longer than exposure time, due to sensory memory read-out effects.

Second, Study 2 aimed at generalizing our results by using three exposure times: 33 ms, 100 ms, and 500 ms. The 33 ms exposure time, in particular, was chosen based on the trustworthiness rating study of Todorov et al. (2009). This study showed that even when exposure time was brought down to 33 ms, participants' trustworthiness ratings were still significantly correlated with the ratings that were given with an unlimited exposure time. Study 2 allowed us to investigate whether accurate trustworthiness detection could be observed at such an extremely short exposure time.

STUDY 2

METHOD

Participants

A total of 75 Belgian university students (36 females, mean age = 21.75, SD = 2.49, range = 29-28) participated voluntarily. Participants were informed that they could receive a monetary compensation depending on their performance in the game (0 euro, 4 euros, or 6 euros).

Material and procedure

Material and procedure were similar to Study 1. Participants were randomly allocated to one of three exposure time groups (33 ms, 100 ms, and 500 ms; see Todorov et al., 2009). After the exposure time had elapsed the Trustee picture was replaced by a masking image that was presented for 160 ms. As in Todorov et al. (2009) the masking image was composed of facial

segments of random faces, which were rearranged to form a jumbled, mosaic image. The masking image was also turned to black-and-white. To familiarize participants with the brief exposure times and mask, two practice rounds with a neutral face were played before the experiment started. We decided against the creation of a cover story to rationalize the short presentation times and masking procedure to participants. Participants were simply instructed that the picture of the Trustee would be "briefly" presented and that it would be followed by a random mosaic pattern.

RESULTS

For each participant we calculated the average transfer rate to cooperators and abusers. These averages were subjected to a 3 (exposure time, between-subjects; 33 ms, 100 ms or 500 ms) x 2 (Trustee strategy, within-subjects; cooperator or abuser) mixed model ANOVA.[2] Results indicated that the main effects of Trustee strategy, $F(1, 72) = 3.61$, $p = .062$, and exposure time, $F(1, 72) = 1.66$, $p = .19$, failed to reach significance. However, the two factors interacted, $F(1, 72) = 10.09$, $p < .001$, $\eta_p^2 = .05$. As Figure 2 (bottom panel) shows, in the 100 ms and 500 ms conditions we replicated the previously observed higher transfer to cooperators than to abusers, $t(49) = 3.7$, $p < .001$, $d = 0.68$, with a 95% confidence interval of 6-22 percentage point for the difference in transfer rate between cooperators and abusers. This pattern was not observed and even reversed in the 33 ms condition, $t(24) = -2.4$, $p = .02$, $d = -0.42$, with a 95% confidence interval of 2-21 percentage point for the difference in transfer rate between abusers and cooperators.

Results were robust to the inclusion of the sex of the trustee as an additional within-subject predictor. The sex of the trustee did not significantly interact with any other factor, but it had a main effect on transfer, $F(1,72) = 18.8$, $p < .001$, $\eta_p^2 = .07$. Female trustees received more transfers (57%) than male trustees (41%). Finally, we computed the sensitivity index d' in

[2] As in Study 1, the analysis does not include transfers to neutral trustees. The 95%-confidence interval for transfers to neutral trustees was 24—68 in the 500 and 100 ms conditions, and 39—81 in the 33 ms condition.

the 500ms, 100ms, and 33ms groups. This index took the values 0.57, 0.20, and -0.31, respectively. The 95% confidence intervals for the difference in *d'* indicate that the 33ms group differed from the two other groups (0.04 — 0.99 for the 100ms group, 0.40 — 1.36 for the 500ms group), whereas it did not differ between the 100ms and 500ms group (-0.11 — 0.84).

In sum, there is no evidence for (and actually evidence against) accurate trustworthiness detection after 33 ms, but renewed evidence for accurate trustworthiness detection after 100 (and 500) ms. To further strengthen our conclusion regarding the null effect of exposure time on trustworthiness detection within the 100-5500 ms window, we conducted two additional analyses.

First, combining the data of Studies 1 and 2, we conducted a 4 (exposure time, between-subjects; 100 ms, 100ms masked, 500 ms masked, 5500 ms) x 2 (Trustee strategy, within-subjects; cooperator or abuser) mixed model ANOVA. Not surprisingly, there was a main effect of Trustee strategy, $F(1, 87) = 17.44$, $p < .0001$, $\eta_p^2 = .17$, with overall 12% more frequent transfer to cooperators (52% transfer) than to abusers (40% transfer). However, neither the main effect of exposure time, $F(3, 87) = 1.64$, $p = .187$, nor the interaction with Trustee strategy, $F(3, 87) = 1.19$, $p = .317$, reached significance. This suggests that varying exposure time in the 5500 ms to 100 ms window has no impact on the accuracy of trustworthiness detection.

Second, we acknowledge that our conclusion that shorter exposure times do not impair trustworthiness detection is based on a null finding. The null-hypothesis significance testing approach that we followed so far does not make it possible to quantify the degree of support for the null hypothesis. An alternative in this case is Bayesian hypothesis testing using Bayes factors (e.g., Masson, 2011; Wagenmakers, 2007). We used the JASP package (JASP Team, 2016) to run a 2 (Trustee strategy) x 4 (exposure time) Bayesian ANOVA with default priors (e.g., Cauchy prior width r = .707) as alternative for our final analysis that focused on the four different exposure time conditions within the 100 ms to 5500 ms window. The full JASP output table is presented in the Electronic Supplementary Material 1. Results showed that the model that received the most support against the Null model was the one with a main effect of Trustee strategy only (BF10 = 577.99). Adding the interaction to the model decreased the degree of support against the Null model (BF10 = 39.56). Hence, the model with a main effect of Trustee

strategy only was preferred to the interaction model by a Bayes factor of 14.61. Following the classification of Wetzels et al. (2011) these data provide strong evidence against the hypothesis that the accuracy of trustworthiness detection is affected by differential exposure time within the 100 ms to 5500 ms range.

GENERAL DISCUSSION

Two studies established that accurate detection of trustworthiness based on the facial features of an unknown Trustee is possible with minimal picture exposure time. The fact that accuracy survived exposure times of 500 ms or 100 ms, together with earlier results establishing that accuracy survived concurrent cognitive load, is conclusive evidence that trustworthiness detection from faces is an intuitive process that does not require deliberate reflection. Importantly, this rules out a possible explanation of the accuracy gap between trustworthiness ratings (i.e., answering a question: *Does this person look trustworthy?*) and trustworthiness decisions (i.e., choosing a course of action: *Do I trust this person with my money?).* Specifically, the current findings rule out the possibility that the greater accuracy of trustworthiness decisions would result from the engagement of reflective processing (whereas trustworthiness ratings would rely on intuitive processing).

Intriguingly, trustworthiness detection disappeared and even reversed with an exposure time of 33 ms, which previous research showed to be sufficient to elicit robust trustworthiness ratings (Todorov et al., 2009). Indeed, participants were even more likely to transfer to abusers than to cooperators in this case. We have no solid explanation for this finding, and one reason to be careful here is that the trust game paradigm might not be entirely suitable for an extremely short exposure time. Even though participants are told that a face appeared on the screen, a 33 ms presentation hovers on the verge of the conscious perception threshold (Axelrod, Bar, & Rees, 2014; Freeman et al., 2014). In a rating study, where participants are simply instructed to give their first impression of a face, a quasi-subliminal presentation may not seem awkward. In a trust game, though, participants are seeking to make money through good decisions. In this context, participants to our study expressed their annoyance at the 33 ms condi-

tion, in which the presentation of the Trustee's picture was barely perceivable. In any case, we must keep in mind that the exact cut-off for accurate trustworthiness detection (33ms or 100 ms) is not instrumental for our main conclusion. That is, even if accurate trustworthiness detection was only possible after 100 ms, this threshold would still largely suffice to characterize it as an intuitive, non-deliberative process.

There are many variants of the Trust Game (e.g., Brülhart & Usunier, 2012) and there are other games which can capture trust behaviors in economic interactions. The decision not to transfer in the Trust Game may be based on other factors than pure distrust; for example, on the desire to guarantee that one is better off than the other player, or on betrayal aversion. Reversely, the decision to transfer in our Trust Game variant may also be affected by inequality aversion (see also Thielman & Hilbig, 2015, for a review of components of trust behavior). To capture these motivations, future studies might opt to use other games or other variants of the Trust Game (e.g., Bohnet & Zeckhauser, 2004; Insko, Kirchner, Pinter, Efaw, & Wildschut, 2005).

The present studies rule out a possible explanation of the accuracy gap between trustworthiness ratings and trustworthiness decisions in the Trust game. However, they do not yet identify the factor that drives the discrepancy. Possible candidates include the fact that Trust game decisions are typically incentivized whereas rating studies are not, for example. Future studies may thus attempt to incentivize trustworthiness ratings. An intriguing (and complementary) possibility is that participants asked for trustworthiness ratings may expect to be asked for justifications next, more so than participants asked for trust decisions. Accordingly, they may base their ratings on invalid yet socially shared and easily communicable cues.

Finally, we emphasize that it is very important not to distort the current findings into any recommendation for people to "trust their guts" or "listen to their intuition" when they form an impression of trustworthiness from the facial features of an interlocutor. This is inadvisable for at least three reasons.

First, here and in previous studies, effect sizes suggest a very limited capacity for trustworthiness detection. Overall, we observed an effect size (difference between the likelihood of transfers to cooperators and the likelihood of transfers to abusers) of about 12 percentage points, with pictures we knew from previous studies to have a good discriminability. With a

random selection of pictures, the effect is closer to 6 percentage points (Bonnefon et al., 2013). This means that while participants trusted better than a random choice algorithm, they frequently transferred to abusers or decided not to transfer to cooperators. Accordingly, decisions made from faces do not necessarily outperform simple rules such as "trust no one" or "trust everyone". In our artificial environment where abusers are encountered as frequently as cooperators, trusting everyone would be a bad strategy (expected payoff of 3.1), but people would still be better off trusting no one (expected payoff of 4) than listening to their intuition (expected payoff of 3.8). Conversely, in a natural environment where most people honor trust (Johnson & Mislin, 2011; Van Lange, 2015), people would quite certainly be better off trusting everyone, than listening to their intuitions about facial features (Todorov, Funk, & Olivola, 2015). Overall, people discriminate between trustees—but the effects are small, and given the base rates of abuse and cooperation, following one's hunches is likely to lead to lower financial earnings in the long run than indiscriminate trust.

Second, over and beyond economic considerations, to encourage people to listen to their intuitions about faces is arguably irresponsible from an ethical perspective. Encouraging people to make decisions about other individuals based on their facial features paves the way to what Olivola et al. (2014) called *face-ism*, that is, unwarranted and consequential decisions based on prejudice about what honest people look like.

Third and last, it is doubtful that the level of accuracy which we observe in controlled laboratory experiments can be sustained in everyday conditions. Bonnefon et al. (2013) observed that trustworthiness detection was only accurate with cropped pictures focused on inner facial features (Figure 1B), and not with full pictures displaying clothing and hairstyle (Figure 1A). We actually attempted to replicate the current Study 1 with full pictures instead of cropped pictures (N= 40, 25 women, mean age = 21.4, SD = 1.93, range = 20-28), and did not observe any accurate trustworthiness detection, neither for a 100 ms exposure nor for a 5500 ms exposure (no effect of Trustee strategy, exposure time, or their interaction, all Fs < 1). The fact that accurate trustworthiness detection seems to require an exclusive focus on inner facial features is one more reason not to encourage people to rely on facial impressions in everyday

life. Because everyday interactions do not take place with cropped faces, facial trustworthiness detection is unlikely to yield good decisions in practice.

In sum, although we must be careful not to give people the impression that they can or should listen to their intuition about the faces of other people, we still need to understand why trustworthiness detection can be accurate in controlled laboratory conditions. Here we showed that accurate trustworthiness detection is possible after only 100 ms exposure to the face of another individual. This finding affords a strong constraint on developing theories of facial trustworthiness detection: Whatever process we theorize to underlie successful trustworthiness detection, needs to be completed in a tenth of a second.

ACKNOWLEDGEMENTS

Thanks to Sem Van Dommelen and Katrien Desmyter for their help in running the studies. We also gratefully acknowledge support through the Institute for Advanced Study in Toulouse (IAST) and ANR-Labex IAST.

DATA SHARING

Raw data can be retrieved from the publishers' website (see Electronic Supplementary Material 2-5) or accessed at osf.io/z86fq

ELECTRONIC SUPPLEMENTARY MATERIAL

ESM 1. Bayesian pooled analysis (ESM 1 Bayesian.pdf)

JASP output file

ESM 2. Raw data Study 1 (raw data study 1.csv)

CSV file with raw data

ESM 3. Raw data Study 2 (raw data study2.csv)

CSV file with raw data

ESM 4. Raw data full picture study (raw data full picture study.csv)

CSV file with raw data full picture study (see General Discussion)

ESM 5. Legend (legend.txt)

Legend for file column headings

REFERENCES

Axelrod, V., Bar, M., & Rees, G. (2014). Exploring the unconscious using faces. *Trends in Cognitive Sciences, 19*, 35-45.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*, 122-142.

Bohnet, Iris, and Richard Zeckhauser. 2004. Trust, risk and betrayal. *Journal of Economic Behavior and Organization, 55,* 467-84.

Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *Journal of Experimental Psychology: General, 142*, 143-150.

Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2015). Face-ism and kernels of truth in facial inferences. *Trends in Cognitive Sciences, 19,* 421-422.

Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2017). Can we detect cooperators by looking at their face?. *Current Directions in Psychological Science*.

Brülhart, M., & Usunier, J. C. (2012). Does the trust game measure trust?. *Economics Letters, 115*, 20-23.

Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., & Seabright, P. (2015). Honest signaling in trust interactions: smiles rated as genuine induce trust and signal higher earnings opportunities. *Evolution and Human Behavior, 36*, 8-16.

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61*, 87-105.

Ciriolo, E. (2007). Inequity aversion and trustees' reciprocity in the trust game. *European Journal of Political Economy, 23*, 1007-1024.

De Neys, W. (2006a). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science, 17*, 428–433.

De Neys, W. (2006b). Automatic-heuristic and executive-analytic processing in reasoning: Chronometric and dual task considerations. *Quarterly Journal of Experimental Psychology, 59*, 1070–1100.

De Neys, W., & Bonnefon, J. F. (2013). The whys and whens of individual differences in thinking biases. *Trends in Cognitive Sciences, 17*, 172-178.

De Neys, W., Hopfensitz, A., & Bonnefon, J. F. (2013). Low second-to-fourth digit ratio predicts indiscriminate social suspicion, not improved trustworthiness detection. *Biology Letters, 9, 20130037*.

De Neys, W., Hopfensitz, A., & Bonnefon, J. F. (2015). Adolescents gradually improve at detecting trustworthiness from the facial features of unknown adults. *Journal of Economic Psychology, 47*, 17-22.

Delhey, J., & Newton, K. (2003). Who trusts? The origins of social trust in seven societies. *European Societies, 5,* 93–137

Efferson, C., Vogt, S. (2013). Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports, 3*, 1-7.

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdale. *Journal of Cognitive Neuroscience, 19*, 1508-1519.

Evans, A. M. & Krueger, J. I. (2009). The psychology (and economics) of trust. *Social and Personality Psychology Compass, 3,* 1003-1017.

Evans, J. St. B. T., & Stanovich, K. (2013). Dual process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*, 223–241.

Freeman, J. B., Stolier, R. M., Ingbretsen, Z. A., & Hehman, E. (2014). Amygdala responsivity to high-level social information from unseen faces. *Journal of Neuroscience, 34*, 10573-10581.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics, 115*, 811-846.

Hertwig, R., & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior, 18*, 59-92.

Insko, C. A., Kirchner, J. L., Pinter, B., Efaw, J., & Wildschut, T. (2005). Interindividual-intergroup discontinuity as a function of trust and categorization: The paradox of expected cooperation. *Journal of Personality and Social Psychology, 88,* 365-385.

JASP Team. (2016). *JASP (Version 0.8) [Computer software].*

Johnson, N., & Mislin, A. (2011). Trust games: a meta-analysis. *Journal of Economic Psychology, 32*, 865-889.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Strauss, Giroux.

Little, A. C., Jones, B. C., DeBruine, L. M., Dunbar, I. M. (2013). Accuracy in discrimination of self-reported cooperators using static facial information. *Personality and Individual Differences, 54*, 507-512.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*, 679-690.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences, 18,* 566-570.

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior, 34*, 83-110.

Rezlescu, C., Duchaine, B., Olivola, C. Y., Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS ONE 7*, e34293.

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*, 409-426.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74*, 1–29.

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science, 21*, 349-354.

Thielmann, I., & Hilbig, B. E. (2015). Trust: An integrative review from a person–situation perspective. *Review of General Psychology, 19*, 249.

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,* 422-423.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*, 1623-1626.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2014). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66,* 519-545.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*, 813-833.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*, 455-460.

Tognetti, A., Berticat, C., Raymond, M., & Faurie, C. (2013). Is cooperativeness readable in static facial features? An inter-cultural approach. *Evolution and Human Behavior, 34*, 427-432.

Tomasello, M. (2009). *Why We Cooperate*. Cambridge, MA: MIT Press.

Van Lange, P. A. M. (2015). Generalized trust: Four lessons from genetics and culture. *Current Directions in Psychological Science, 24*, 71-76.

van ‘t Wout, M., & Sanfey, A. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*, 796-803.

Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*, 779-804.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t-tests. *Perspectives on Psychological Science, 6*, 291-298.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after 100 ms exposure to a face. *Psychological Science, 17*, 592-598.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal sentencing outcomes. *Psychological Science, 26,* 1325-1331.

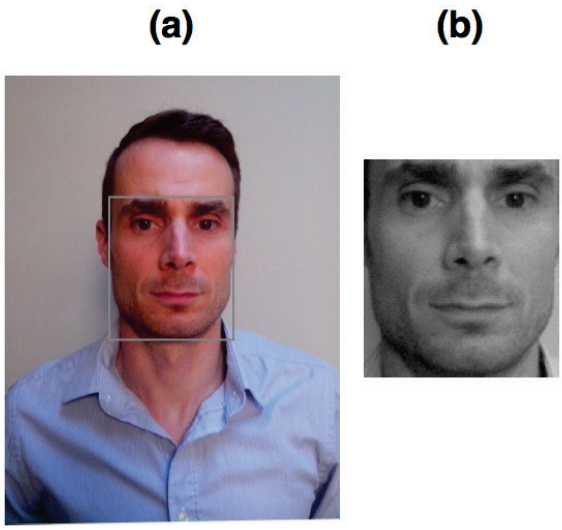

Figure 1. Mock example of the experimental material. (a) Original full picture. (b) Cropped, resized black and white version.

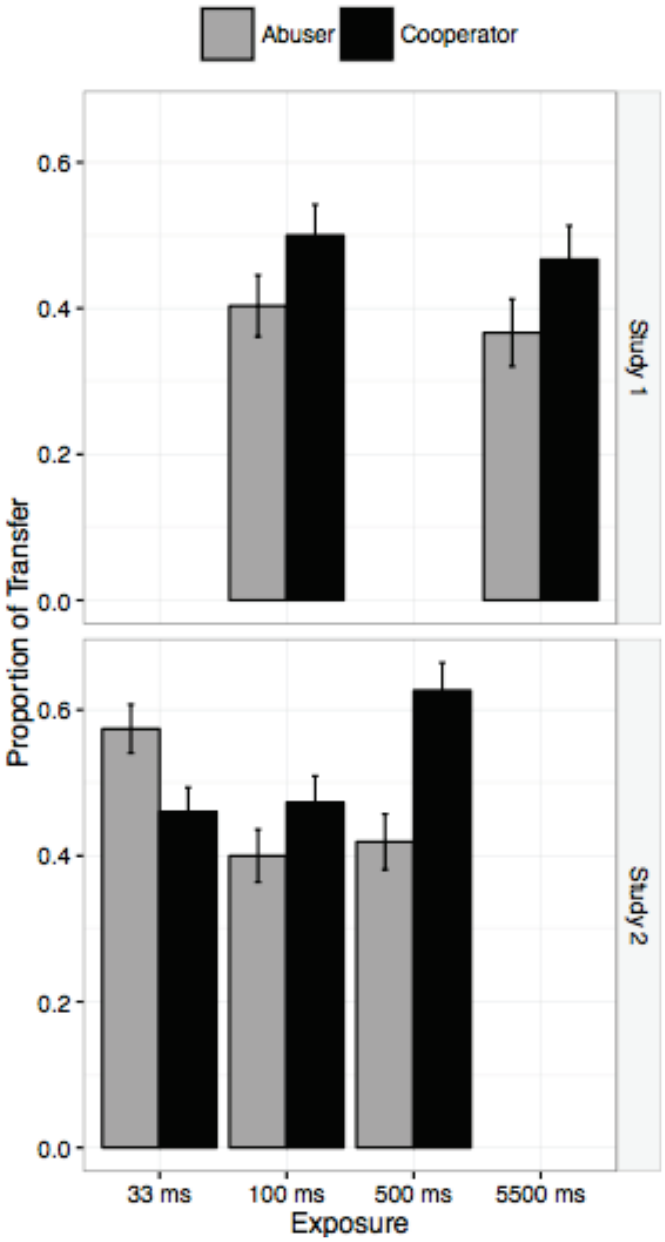

Figure 2. Proportion of transfer to Cooperators and Abusers as a function of picture exposure time in Study 1 (top panel, without masking), and Study 2 (bottom panel, with masking). Error bars show the standard error of the mean.