TSE working paper series

12-311

Research Group: Behavioral and Experimental Economics

May, 2012

"The Modular Nature of Trustworthiness Detection"

Astrid Hopfensitz,
Jean-François Bonnefon
and
Wim De Neys

Toulouse School of Economics (TSE) - Manufacture des Tabacs
Aile Jean-Jacques Laffont - 21, allée de Brienne - 31015 TOULOUSE Cedex 6
Tél : +33 (0) 5 61 12 85 89 - Fax : +33 (0) 5 61 12 86 37 - **www.tse-fr.eu** - contact@tse-fr.eu

Toulouse School of Economics

Look for the final version in JEP: General

The Modular Nature of Trustworthiness Detection

Jean-François Bonnefon
Centre National de la Recherche Scientifique

Astrid Hopfensitz
Toulouse School of Economics

Wim De Neys
Centre National de la Recherche Scientifique

The capacity to trust wisely is a critical facilitator of success and prosperity, and it has been conjectured that people of higher intelligence were better able to detect signs of untrustworthiness from potential partners. In contrast, this article reports five Trust Game studies suggesting that reading trustworthiness off the faces of strangers is a modular process. Trustworthiness detection from faces is independent of general intelligence (Study 1) and effortless (Study 2). Pictures that include non-facial features such as hair and clothing impair trustworthiness detection (Study 3) by increasing reliance on conscious judgments (Study 4), but people largely prefer to make decisions from this sort of pictures (Study 5). In sum, trustworthiness detection in an economic interaction is a genuine and effortless ability, possessed in equal amount by people of all cognitive capacities, but whose impenetrability leads to inaccurate conscious judgments and inappropriate informational preferences.

Trust is a critical facilitator of cooperation, and the cornerstone of prosperous societies (Zak & Knack, 2001). The problem with trust, though, is that it cannot be given out indiscriminately, for it can easily be abused. An important challenge for the social and cognitive sciences is thus to lay bare the processes that allow people to trust wisely. Accordingly, the issues of trust, trustworthiness and trustworthiness detection have garnered growing attention from multiple fields within and beyond psychology. One view that is currently gaining momentum is the *Intelligence-Trust conjecture* (Sturgis, Read, & Allum, 2010), or the assumption that smarter individuals are better at detecting signs of untrustworthiness in social and economic situations. It was found that the general propensity to trust correlates with intelligence (Sturgis et al., 2010; Schoon & Cheng, 2011; Segal & Hershberger, 1999), with a greater ability at trusting wisely (Yamagishi, Kikuchi, & Kosugi, 1999), and with life achievements (Delhey & Newton, 2003; Paxton, 2007). The Intelligence-Trust conjecture makes sense of this array of findings by assuming that smarter individuals do not necessarily start trusting more than others, but experience greater success with their trust decisions. Being better equipped to detect signs of untrustworthiness in potential interaction partners, they would allocate their trust wisely and experience positive reinforcement about trusting others. Over the course of a lifetime, they would accordingly end up being more trusting and more prosperous.

The Intelligence-Trust conjecture has never been subjected to direct and rigorous experimental testing (Sturgis et al., 2010; Evans & Krueger, 2010), and it could appear at odds with results derived from the evolutionary approach to social exchange (Cosmides, Barrett, & Tooby, 2010). It has been suggested that the human mind has tackled the evolutionary problem of social exchange by developing a specialized module for cheater detection, which does not rely on central cognitive processing. Cheater detection is thought to be unrelated to general intelligence, undemanding in cognitive resources, and impenetrable to consciousness. It would seem plausible that the mind developed broadly similar cognitive architectures for detecting trustworthiness and for detecting cheaters (Verplaetse, Vanneste, & Braeckman, 2007). As an implication, and pace the Intelligence-Trust conjecture, trustworthiness detection (TD) would be fairly routinized and independent of general intelligence. The goal of this article is to explore this hypothesis. Our paradigm for measuring TD is the classic Trust Game (Berg, Dickhaut, & McCabe, 1995). In this game a player (the Investor) is endowed with an amount of money and decides whether she will transfer this endowment to another player (the Trustee). If the endowment is transferred, it is multiplied by a factor, and the Trustee then decides how much to send back to the Investor. In a sense, the Investor is betting on the trustworthiness of the Trustee: A perfectly accurate performance at TD would allow Investors to transfer to those and only those Trustees whose decision is to reciprocate.

Recent research on TD in the trust game has focused on the signals that Investors pick from the Trustees' facial features. Faces are rapidly appraised for trustworthiness (Yang, Qi, Ding, & Song, 2011) and this initial appraisal strongly influences subsequent decisions in the trust game, in the

Address correspondence to Jean-François Bonnefon, 'Cognition, Langues, Langage et Ergonomie,' Maison de la recherche, 5 allées A. Machado, 31058 Toulouse Cedex 9, France. E-mail: bonnefon@univ-tlse2.fr. This work was partially supported by grant PHC TOURNESOL FL 2011 21798SH, grant ANR 2010 JCJC 1803 01, the MSHS-Toulouse, and the Swiss & Global-Ca' Foscari Fondation.

short term (van't Wout & Sanfey, 2008) and in the long term (Chang, Doll, van't Wout, Frank, & Sanfey, 2010). Most interestingly, the face might offer valid cues to trustworthiness. For example, people can accurately discriminate cooperators from non-cooperators based on a picture of their face at the moment they were pondering whether to cooperate (Verplaetse et al., 2007; Willis & Todorov, 2006). Furthermore, men with greater facial width are perceived as less trustworthy, and it turns out that they are less likely to honor trust in the trust game (Stirrat & Perrett, 2010). This prior evidence suggests that people can read trustworthiness off the face of their economic partners. The current article offers a series of experiments that triangulate the cognitive nature of TD, seeking evidence for its automaticity and encapsulation. Although not all evolved modules possess these two features (Barrett, Frederick, Haselton, & Kurzban, 2006), their presence is a strong indication that central processing is not involved. Note that even if TD turns out to be a module it could still be used for the detection of a range of personality traits besides trustworthiness, an issue to which we will get back in the final section of this article.

If TD is automatic and encapsulated then Investors of higher general intelligence should have no advantage at TD; TD should be impervious to concurrent cognitive load; accurate TD should be decoupled from conscious trustworthiness judgments; and Investors should be oblivious to the boundary conditions of accurate TD. Finding evidence for automaticity and encapsulation would put us in a strong position to claim that accurate TD is a modular rather than central process.

Study 1: Intelligence and trustworthiness detection

To investigate whether more intelligent Investors would be better at TD, we took the pictures of 60 Trustees and recorded their strategies. We could thus identify 35 'reciprocators' (returning more than what they were transfered), 7 'abusers' (returning zero), and 18 neutral Trustees (returning the exact amount that was transfered). Trustees' strategies and pictures were taken from a previous study (Centorrino, Djemai, Hopfensitz, Milinski, & Seabright, 2011) in which 79 young adults (aged between 18 and 35 years) were familiarized with the Trust Game and asked to play the role of Trustee. They were informed that the Investor would be endowed with an amount of money which was multiplied by the factor 3 in case the Investor decided to transfer the endowment. Participants were asked to indicate how much they would send back in case the Investor transferred the endowment. They were given three options: return zero euro, return the exact amount that was transferred, or return half of the new global amount. All trustees were clearly informed that they would be randomly paired with one Investor and receive the money they made based on their strategy. Pictures were extracted from movies that were recorded of Trustees after they had been informed about the general structure of the game.

For the present paper we selected 60 pictures (30 male and 30 female) from the original 79 Trustees. Abusers and reciprocators were naturally matched for age, as all Trustees were young adults recruited on campus. The 60 pictures were selected so that the proportion of abusers and reciprocators would be similar for male Trustees (4 abusers, 18 reciprocators) and female Trustees (3 abusers, 17 reciprocators). To increase homogeneity in our pictures set, we avoided pictures of distinctively non-Caucasian Trustees, and pictures of Trustees whose facial expression was not neutral enough.

Method

A total of 208 undergraduates from the University of Leuven (Belgium) were familiarized with the Trust Game and asked to play the role of Investor. They played 60 single shot games, each time with a different Trustee. Participants were endowed with 4 euro on each game. Each game started with a fixation cross that was presented for 1000 ms. Next, the picture of the Trustee was presented for 5500 ms. As in previous studies, we presented black-and-white pictures of the Trustees' faces that were cropped to minimize any display of clothing or hairstyle. The horizontal cropping points were set at the left and right facial boundary. Vertical cropping points were the chin and top of the eyebrows, respectively (Figure 1A). After the presentation of the picture participants were asked whether they wanted to transfer their money to the Trustee or not by pressing 1 (transfer) or 2 (no transfer). After participants entered their response they were asked to press the space bar whenever they were ready to start the next game. Participants did not receive feedback about their decisions after each individual game. Participants were informed that after the experiment one game would be randomly selected and they would receive whatever money they made in that game. Participants' fluid Intelligence score was measured by means of the short version (14 items) of Raven's advanced progressive matrices (Bors & Stokes, 1998). The test took about 20 minutes and participants completed it after they finished the Trust Game. Scores range from 0 to 12 as the first two items are conceived as practice trials.

Results

Trustworthiness detection. Investors transfered money to 41% of the abusers (SE = 1.4) and 46% of the reciprocators (SE = 1.7), an effect size $h = 0.10$. Figure 2A displays the average transfer rates to reciprocators and abusers as a function of the Raven score of the Trustees (bottom half and top half of the distribution). We ran an analysis of variance with Tranfer rate as the dependent variable, Trustee's strategy as a within-subject factor, and Raven score as a covariate. The analysis detected a main effect of Trustee's strategy, $F(1, 206) = 13.3$, $p < .001$. This effect was not moderated by the Raven score of the Investors, $F(1, 206) < 1$, $p = .73$, which itself did not have any detectable main effect, $F(1, 206) < 1$, $p = .60$.[1] Study 1 thus demonstrates that Investors were able to detect trustworthiness: Investors were more likely to transfer to reciprocators than to abusers.

[1] Raven scores spanned the full 0–12 range, $M = 7.03$, SD = 2.55. This distribution was similar to that observed in the Bors and Stokes (1998) norming study for first-year undergraduates.

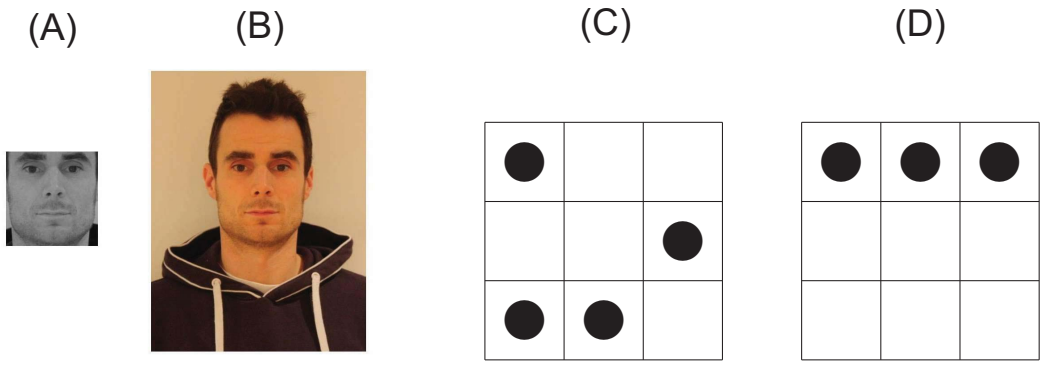

Figure 1. Examples of experimental materials. (A) Cropped picture of a Trustee. (B) Full picture of the same Trustee. (C) A dot pattern used in the Load condition. (D) A dot pattern used in the Control condition.

In line with our modular hypothesis, this ability was unrelated with a standard measure of intelligence. Individuals with lower Raven scores were as capable of detecting trustworthiness as individuals with higher Raven scores.

Additional results. We ran a supplementary analysis of variance on Transfer rates in which we introduced Trustee's gender as an additional within-subject factor. Results were robust to the introduction of this control variable. The analysis again detected a main effect of Trustee's Strategy, $F(1, 206) = 15.3$, $p < .001$; no main effect of the Raven score, $F(1, 206) < 1$, $p = .61$; and no interaction between Trustee's strategy and Raven score, $F(1, 206) < 1$, $p = .69$. The analysis detected two additional effects: a main effect of Trustee's gender (more transfer to female Trustees), $F(1, 206) = 9.9$, $p = .002$, qualified by an interaction with Trustee's strategy, $F(1, 206) = 19.0$, $p < .001$, reflecting greater trustworthiness detection for female faces in our sample. We also ran an analysis of variance with response speed as the dependent variable, Trustee's strategy as a within-subject factor, and Raven score as a covariate. This analysis indicated that Investors were faster when presented with pictures of abusers (1.28 faces per second, SE = 0.05), than when presented with pictures of reciprocators (1.00 face per second, SE = 0.03). This difference was statistically significant, $F(1, 206) = 35.6$, $p < .001$. It was not moderated by the Raven score, $F(1, 206) < 1$, $p = .40$, which itself did not have any detectable effect on response speed, $F(1, 206) = 2.1$, $p = .15$.

Study 2: Trustworthiness detection and concurrent cognitive load

In a second experiment, we tested whether TD was impeded by cognitive load: if TD does not involve central executive processing, then it should remain accurate even when Investors are cognitively burdened by a concurrent task. To test this hypothesis, we ran a replication of the first study on an independent sample of 93 Investors, in which we introduced a cognitive load manipulation.

Method

A total of 93 undergraduates from the University of Leuven (Belgium) played the same Trust Game as in Study 1, under concurrent cognitive load. Before the picture of the Trustee was shown, a dot pattern in a 3×3 matrix was flashed for 900 ms. Participants had to keep the pattern in memory while they saw the picture and made their transfer decision. After participants had entered their transfer response they were presented with an empty matrix and had to indicate the location of the dots. Participants were randomly assigned to the Load and Control group. In the Load group the matrix was filled with a complex 4-dot pattern (see Figure 1C for an example), whose storage efficiently taps executive resources (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001; De Neys, 2006). In the control group the pattern consisted of three dots on a horizontal or vertical line (see Figure 1D for an example), whose storage places but a minimal burden

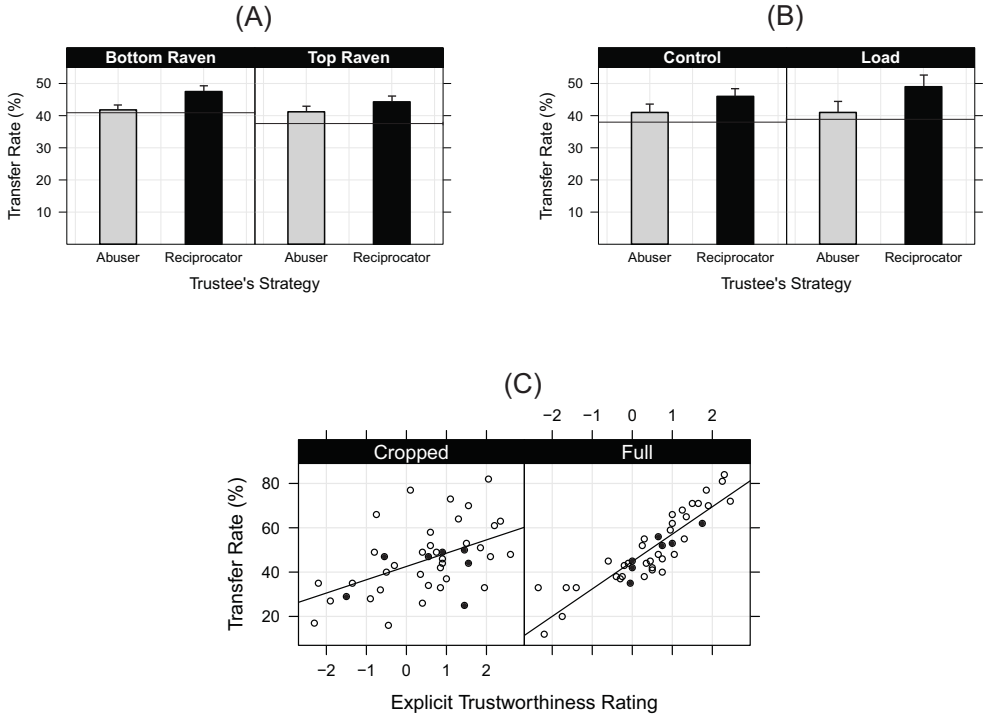

Figure 2. Main experimental results. Error bars show standard error of the mean. The horizontal lines in A and B display the transfer rates to neutral trustees, who send back the exact amount they were transferred. (A) Investors transfer more to Trustees who privately decided to reciprocate, than to Trustees who privately decided to abuse, and this capacity to detect trustworthiness is independent of Investors' Raven scores of fluid intelligence. (B) Investors transfer more to Trustees who privately decided to reciprocate, than to Trustees who privately decided to abuse, and this capacity to detect trustworthiness is resistant to concurrent cognitive load. (C) Transfer decisions are almost exclusively predicted by explicit ratings of trustworthiness in the full picture trust game (in which trustworthiness detection fails), much less so in the cropped picture trust game (in which trustworthiness detection is more successful). Abusers are represented by black dots.

on executive resources. Participants were familiarized with the dot memorization task before the Trust Game started, on two practice trials. Training instructions stressed that it was crucial that the dot pattern was reproduced correctly in the upcoming task. After each individual Trust Game trial participants also got feedback about their memorization performance.

Results

Dot matrix task. The concurrent memorization task was properly performed. The mean number of correctly localized dots for the complex dot pattern was 3.74 (out of 4, SE = 0.03) and 2.91 (out of 3, SE = 0.01) for the simple pattern in the control condition. Thus, overall, about 94% of complex patterns and 97% of simple patterns were reproduced correctly. These high accuracy rates confirm that participants did as instructed, and engaged executive resources into the memorization task.

Trustworthiness detection. Overall, Investors transfered money to 41% of the abusers (SE = 2.6) and 48% of the reciprocators (SE = 2.1), an effect size $h = 0.14$. As shown in Figure 2B, the difference in transfer rates was present both in the control condition and under complex concurrent load. We ran an analysis of variance with Tranfer rate as the dependent variable, Trustee's strategy as a within-subject factor, and Load as a between-subject factor. The analysis detected a main effect of Trustee's strategy, $F(1,91) = 14.3$, $p < .001$. This effect was not moderated by concurrent load, $F(1,91) < 1$, $p = .60$, which itself did not have any detectable main effect, $F(1,91) < 1$, $p = .52$. Study 2 thus confirms that Investors are able to detect trustworthiness, and that this capacity is robust to concurrent cognitive load.

Additional results. We ran a supplementary analysis of variance on Transfer rates in which we introduced Trustee's gender as an additional within-subject factor. Results were robust to the introduction of this control variable. The analysis again detected a main effect of Trustee's Strategy, $F(1,91) = 15.2$, $p < .001$; no main effect of load, $F(1,91) < 1$, $p = .46$; and no interaction between Trustee's strategy and concurrent load, $F(1,91) < 1$, $p = .76$. There was no main effect of Trustee's gender, $F(1,91) = 1.6$, $p = .21$, but the analysis detected an interaction between Trustee's gender and Trustee's strategy, $F(1,91) = 16.5$, $p < .001$, reflecting greater trustworthiness detection for female faces in our sample. An analysis of variance conducted on response speed (as a function of Trustee's strategy and load group) in-

dicated that Investors were faster when presented with pictures of abusers (1.04 faces per second, SE = 0.05), than when presented with pictures of reciprocators (0.94 face per second, SE = 0.04). This difference was statistically significant, $F(1, 91) = 4.6$, $p < .05$. It was not moderated by cognitive load, $F(1, 91) < 1$, $p = .37$, which itself did not have any detectable effect on response speed, $F(1, 91) < 1$, $p = .81$.

Study 3: Extension to real-life pictures

In line with previous research (Stirrat & Perrett, 2010), our first and second studies used cropped, black-and-white and standardized for size versions of Trustees' pictures, which did not reveal clothing or hairstyle (Figure 1A). Colored pictures displaying clothing and hairstyle provide additional information about Trustees. To the best of our knowledge, there is no prior evidence that this information is useful, irrelevant or misleading when assessing trustworthiness. However, and in any case, it is possible that the link between intelligence and TD might be weakened when little information is provided, and detectable when more information is provided, giving a role after all to intelligence in TD. In order to control for that possibility, we ran a replication of our first study with an independent sample of 180 Investors, in which we used the unmodified colored pictures of Trustees that showed clothing and hairstyle (Figure 1B).

Method

A total of 180 undergraduates from the University of Leuven (Belgium) played the same Trust Game as in Studies 1 and 2, with the original non-cropped color pictures that we took from our sample of Trustees. As illustrated in Fig. 2B, these full pictures displayed non-facial features such as clothing and hairstyle. Participants completed the same short version of Raven's advanced progressive matrices as in Study 1. After the experiment we also asked participants to write down what percentage of Trustees they generally expected to abuse trust.

Results

Trustworthiness detection. Overall, Investors transfered money to 49% of the abusers (SE = 1.5) and 51% of the reciprocators (SE = 1.9), an effect size $h = 0.04$. Contrary to what we observed in the first two studies, this difference was not statistically significant. We ran an analysis of variance with Tranfer rate as the dependent variable, Trustee's strategy as a within-subject factor, and Raven score as a covariate. There was no effect of Trustee's strategy, $F(1, 178) = 1.9$, $p = .17$, no effect of the Raven score $F(1, 178) = 1.3$, $p = .25$,[2] and no significant interaction, $F(1, 178) < 1$, $p = .56$. This result suggests that TD was impaired when Investors were presented with full pictures. In order to further investigate this possibility, we conducted an additional analysis comparing TD in Studies 1 and 2 (cropped pictures) and TD in Study 3 (full pictures). Running such a between-study comparison was deemed appropriate, since the sampled population, protocols and experimenters were the same across studies.

Between-study comparison. The ANOVA (Transfer Rate as the dependent variable, Trustee's strategy as a 2-level within-participant predictor, Study as a 3-level between-participant predictor) detected a significant effect of Strategy, $F(1, 492) = 25.9$, $p < .001$, and the expected interaction with Study $F(2, 492) = 3.3$, $p = .04$. The analysis also detected a significant effect of Study, reflecting the fact that participants transferred more in Study 3, $F(2, 492) = 4.6$, $p = .01$.[3]

Additional results. We ran a supplementary analysis of variance on Transfer rates in which we introduced Trustee's gender as an additional within-subject factor. Results were robust to the introduction of this control variable. The analysis did not detect any effect of Trustee's Strategy, $F(1, 178) = 1.1$, $p = .29$; no main effect of the Raven score, $F(1, 178) = 1.4$, $p = .23$; and no interaction between Trustee's strategy and Raven score, $F(1, 178) < 1$, $p = .49$. In line with what we already observed in studies 1 and 2, the analysis detected two additional effects: a main effect of Trustee's gender (more transfer to female Trustees), $F(1, 178) = 54$, $p < .001$, qualified by an interaction with Trustee's strategy, $F(1, 178) = 19.0$, $p < .001$, reflecting better TD for female faces in our sample, $F(1, 178) = 4.8$, $p = .03$. We also ran an analysis of variance with response speed as the dependent variable, Trustee's strategy as a within-subject factor, and Raven score as a covariate. Interestingly, this analysis indicated that Investors were slower to make their decisions when presented with pictures of abusers (1.16 faces per second, SE = 0.05), than when presented with pictures of reciprocators (1.32 faces per second, SE = 0.04). This difference was statistically significant, $F(1, 178) = 8.7$, $p = .004$. It was not moderated by the Raven score, $F(1, 178) = 3$, $p = .08$, which itself did not have any detectable effect on response speed, $F(1, 178) < 1$, $p = .33$. The interaction term that comes close to significance reflects a slight tendency for high-Raven subjects to be faster for reciprocators' pictures and slower for abuser's pictures, but the corresponding correlation coefficients are very low (.02 and −.11, respectively). The significant main response speed effect suggests that Investors might retain a gut feeling that some of the Trustees should not be trusted, witness their slower decisions for faces of would-be abusers. But clearly this feeling does not result on a decision not to trust. Finally, note that in Study 3 we also explicitly asked participants to write down what percentage

[2] Raven scores spanned the full 0–12 range, $M = 7.05$, SD = 2.57. This distribution was similar to that observed in the Bors and Stokes (1998) norming study for first-year undergraduates.

[3] Another option to test the effect of picture type in our studies is to run an ANOVA with Transfer Rate as the dependent variable, Trustee's Strategy as a 2-level within-participant predictor, and Picture Type as a 2-level between-participant predictor. The results of this analysis are similar to that we report in the main text.

of Trustees they generally expected to abuse trust. We observed that subjects with higher Raven scores expected less abuse ($r = -.21$, $p = .005$), with the top-quartile expecting 36% abuse, and the lower quartile expecting 47% abuse. This confirms that more cognitively capable subjects, just as in previous research (Sturgis et al., 2010), had a greater propensity to think of others as trustworthy. However, the key point is that this propensity did not translate into behavior, as more cognitively capable subjects did not transfer more than other subjects. Note that this finding helps to allay potential concerns about our measure of intelligence: Arguably, if the 12-item Raven score captures the effect of intelligence on trust attitudes, it should be able to capture the effect of intelligence on TD.

Study 4: Explicit trustworthiness judgments

Our next test focuses on the decoupling of TD and explicit trustworthiness judgments. In line with our modular hypothesis, we suspect that explicit judgments about whether a Trustee is trustworthy (as compared to decisions to transfer money to that Trustee) might not be a good predictor of the Trustee's strategy. That is, if TD is an encapsulated process in which Investors have no insight, their trusting decisions might be more accurate than their explicit trustworthiness judgments.

Method

A total of 80 undergraduates from the University of Leuven (Belgium) were asked to rate how trustworthy each Trustee looked on a 7-point rating scale (-3 to +3). The exact same pictures that were presented in our Trust Games were presented to the participants. Each rating trial started with a fixation cross that was presented for 1000 ms. Next, the picture was presented and participants wrote down their rating on a scoring sheet. In addition to the trustworthiness rating we also asked the participants to rate how intelligent, attractive, and aggressive the Trustee looked. These additional ratings were not intended to be used for the present study but we note here that abusers and reciprocators did not significantly differ on any of these ratings. When participants were finished rating a picture they pressed the space bar and the next trial started. Half of the participants rated the cropped pictures that were used in Studies 1 and 2, the other half rated the full pictures that were used in Study 3. Each participant rated a total of 30 pseudo-randomly selected faces. The randomization procedure guaranteed that each picture was rated by exactly 40 participants.

Results

Trustworthiness detection. In line with our hypothesis, abusers and reciprocators were rated as equally trustworthy-looking, for cropped pictures as well as full pictures. If anything, abusers obtained slightly greater average ratings of trustworthiness than reciprocators, for cropped pictures ($M_{ab} = 0.55$, $M_{re} = 0.48$) as well as full pictures ($M_{ab} = 0.59$, $M_{re} = 0.44$). We ran an analysis of variance on pictures, with trustworthiness ratings as the dependent variable, Trustee's strategy as a between-item factor, and Picture type as a repeated factor. This analysis did not detect any significant main effect or interaction effect, all $F < 1$, all $p > .80$. Introducing Trustee's gender as an additional between-item variable in the analysis did not impact results, but yielded a significant interaction between Picture Type and Trustee's gender, $F(1, 38) = 6.2$, $p = .02$. This result would appear to reflect the fact that females looked more trustworthy than males from full pictures ($M_f = 0.94$, $M_m = 0.03$, $t(40) = 2.73$, $p = .01$), whereas females and males looked equally trustworthy from cropped pictures ($M_f = 0.58$, $M_m = 0.41$, $t(40) = 0.44$, $p = .66$).

Additional results. An additional analysis helps to understand why TD was more accurate with cropped pictures than with full pictures. As manifest in Figure 2C, explicit trustworthiness ratings provided about the full pictures predicted 80% of the variance of transfer decisions in Study 3 ($r = .90$, $p < .001$), whereas explicit trustworthiness ratings provided about the cropped pictures predicted only 18% of the variance of transfer decisions in Study 1 ($r = .43$, $p = .005$). This suggests that subjects playing with the full pictures largely relied on explicit trustworthiness judgments, for impaired performance, whereas subjects playing with cropped pictures did not rely as much on explicit judgments, for better performance.[4]

Study 5: Inappropriate preferences

Investors in the first study (who made decisions from cropped, black-and-white pictures), detected trustworthiness better than Investors in the third study, who saw full (and colored) pictures. If TD is an encapsulated process in which Investors have no conscious insight, they might not be able to realize that cropped, black-and-white pictures improve their decisions, and they might manifest an inappropriate preference for playing the game with full, colored pictures.

Method

A total of 67 voluntary undergraduates from University College Sint-Lieven (Aalst, Belgium) were familiarized with the Trust Game. They were shown two examples of the cropped pictures and two examples of the original full pictures of the same Trustees. The two cropped and two full

[4] Taken together, the findings of Studies 3 and 4 raise an intriguing possibility, evoked by a reviewer. Investors in Study 3 were slower to make their decisions when presented with the pictures of abusers, as if they retained a gut feeling about their trustworthiness but eventually went for a more explicit judgment. Because they had 5500 ms to make a decision, they had ample time to search for more cues in hairstyle and clothing. With a higher time pressure, participants might have been more likely to follow their initial gut feeling based on facial features. If this interpretation is correct, then time pressure should improve trustworthiness detection from full pictures.

pictures were presented side by side on a single slide. We randomly selected four Trustees from our group of reciprocators and four Trustees from our group of abusers. Two sets of slides were created with each slide showing the picture of one reciprocator and one abuser. Location of the cropped and full version (right or left hand side of the slide) was counterbalanced across the sets. Half of the participants were presented with Set 1, the other half with Set 2. Participants were asked to write down whether they would prefer to play the Trust Game with the pictures on the left or right hand side.

Results

Preferences did not differ across the two sets (p = .69). In line with our hypothesis, 78% of Investors indicated that they would prefer to play the game with full pictures (binomial, $p < .001$). In sum, Investors did not seem to realize that they would make more accurate transfer decisions from cropped pictures. If Investors had conscious insight in TD processes, they could have been expected to pick the type of pictures that allowed them to make better decisions. The fact that they picked the type of pictures that impairs the quality of their decisions thus suggest that they had no conscious insight in TD processes.

Discussion

The present studies established that people of all cognitive capacities were equally capable of reading trustworthiness off a stranger's face (Study 1), and that this capacity was robust to concurrent cognitive load (Study 2). This capacity was impaired, however, by the presence of external features such as hairstyle and clothing, and higher cognitive capacity did not protect from this detrimental effect (Study 3). When these external features were present, trusting decisions were almost perfectly predicted by explicit trustworthiness judgments (Study 4), for degraded performance. People did not seem to be aware of this degraded performance, as they largely preferred to rely on pictures that showed these external features (Study 5).

Our five studies provide convergent evidence that trustworthiness detection (TD) from faces is an encapsulated, automatic process, in which people have little conscious insight, and for which higher intelligence provides little benefit.[5] The Intelligence-Trust conjecture assumes that smarter individuals do not necessarily start trusting more than others, but experience greater reinforcement through better trust decisions. The current results suggest to look for another explanation of the positive association between higher intelligence and the greater propensity to trust, since accurate TD appears to be a matter of gut feelings rather than reflective engagement. Maybe the next simplest explanation is that more intelligent subjects have early incentives to try out trusting others, and that trust is a self-reinforcing attitude: Few people, at least in laboratory studies such as ours, abuse the trust they have been endowed with. If the base rate of trust exploitation is indeed low, any early factor (including parental encouragement) that encourages to trust more will build up on the long term into more trusting attitudes.

Another possibility is that intelligence impacts other aspects of TD than facial scrutiny: people of higher intelligence would have no advantage at reading trustworthiness off faces, but would be able to use other cues better than people of lower intelligence. Although this possibility cannot be ruled out, its explanatory power is limited by the enduring strength of first impressions. Even in a courtroom environment, which put strong emphasis on the reflective evaluation of credibility and trustworthiness, jurors make fast and strong initial judgments based on a defendant's face, which influence the manner in which they process subsequent information (Porter, Gustaw, & ten Brinke, 2010). If first impressions have a powerful and lasting influence on trust, and if intelligence does not increase the accuracy of first impressions, then the overall impact of intelligence on accurate TD is likely to be limited.

The current results do tell us that TD from faces is a genuine ability, though, and shed light on its strengths and boundary conditions. As an automatic process, TD appears to be relatively effortless, and independent of one's cognitive capacities. It is, however, quite fallible, witness the small effect sizes obtained in Studies 1 and 2. Even though participants could significantly detect trustworthiness, their performance was not as good as that observed, for example, in the Verplaetse et al. (2007) study where pictures were taken at the moment the targets made their decisions to cooperate or not. This suggests that the facial expression related to the decision to cooperate could provide a more powerful cue to trustworthiness than the face as such, and we cannot rule out yet that intelligence might play a role in the processing of this cue. Beyond facial expression, other verbal and behavioral cues might be factored in everyday TD (or lie detection) situations, and our data are silent as to whether some of these cues might be processed centrally rather than in a modular fashion.

Not only is TD fallible, but it seems easy to manipulate. This should not be a surprise, considering that Ponzi schemers and other fraudsters make a living from failures of TD. Bernard Madoff, who operated the largest documented Ponzi scheme in history, was able to convince hundreds of persons to trust him with their savings. Just as other successful fraudsters, his scheme relied in part on his being perceived as sincere, ingratiating and trustworthy. The Madoff case, and similar if less publicized cases, should serve as a warning not to adopt too rosy a view of the human ability at TD.

In our studies, TD was overridden by external features (e.g., hair), which people preferentially attended to, to the detriment of the internal facial features that would have served TD best. Although external facial features are known to improve the detection of traits such as extraversion and

[5] That is not to say that other factors might not moderate TD, be they forms of intelligence not captured by the Raven test, or other individual differences factors. In light of the lively debate about whether some people have a better ability to spot liars than others (O'Sullivan & Ekman, 2004), we should be careful not to hastily rule out the possibility that some people might be better at TD than others.

physical health (Kramer & Ward, 2010), and thus carry useful information in many situations, they appear to hurt TD unbeknownst to trusters. This is an interesting element regarding the modularity of TD, since it makes it less likely that the processes underlying TD would reflect the operation of a general system for personality detection. That is not to say, however, that the processes underlying TD cannot be used for (or indeed, were not derived from) the detection of other traits or dispositions, such as aggressiveness or dominance. One challenge for future research will be to map the cluster of dispositions that can and cannot be detected by the same processes used for TD.

Finally, it is important to note that accurate TD was only manifest in our studies in the *decisions* that people made, and not in the explicit judgments that they expressed. This asymmetry can help to explain why accurate TD from faces was not always demonstrated in prior research. For example, in one previous study (Porter, England, Juodis, ten Brinke, & Wilson, 2008), subjects gave explicit judgments of trustworthiness based on faces that included some of America's most wanted criminals. Subjects were hardly capable of detecting the untrustworthy character of these criminals from their faces. Our data suggest that they might have been more successful, if their task had been to make a decision about whether they would trust these individuals with their money or their safety. In sum, the present data suggest that the ability to detect trustworthiness from faces is automatic and encapsulated: it is real, effortless, and manifest in our decisions; but it is fallible, impenetrable, and decoupled from our conscious judgments. Future research can build on this cognitive triangulation of TD processes to lay bare its developmental trajectory, its social moderators, its neural bases, and its applied consequences.

References

Barrett, H. C., Frederick, D. A., Haselton, M. G., & Kurzban, R. (2006). Can manipulations of cognitive load be used to test evolutionary hypotheses? *Journal of Personality and Social Psychology*, *91*, 513–518.

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

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, *58*, 382–398.

Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., & Seabright, P. (2011). Smiling is a costly signal of cooperation opportunities: Experimental evidence from a trust game. *IDEI working paper nr. 669*.

Chang, L. G., 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.

Cosmides, L., Barrett, H. C., & Tooby, J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy of Science of the USA*, *107*, 9007–9014.

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

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

Evans, A., & Krueger, J. (2010). Elements of trust: Risk and perspective-taking. *Journal of Experimental Social Psychology*, *47*, 171–177.

Kramer, R. S. S., & Ward, R. (2010). Internal facial features are signals of personality and health. *Quatrerly Journal of Experimental Psychology*, *63*, 2273–2287.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*, 621–640.

O'Sullivan, M., & Ekman, P. (2004). The wizards of deception detection. In P. Granhag & L. Strömwall (Eds.), (pp. 269–286). Cambridge, MA: Cambridge University Press.

Paxton, P. (2007). Association memberships and generalized trust: A multilevel model across 31 countries. *Social Forces*, *86*, 47–76.

Porter, S., England, L., Juodis, M., ten Brinke, L., & Wilson, K. (2008). Is the face a window to the soul? Investigation of the accuracy of intuitive judgments of the trustworthiness of human faces. *Canadian Journal of Behavioural Science*, *40*, 171–177.

Porter, S., Gustaw, C., & ten Brinke, L. (2010). Dangerous decisions: The impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime and Law*, *16*, 477–491.

Schoon, I., & Cheng, H. (2011). Determinants of political trust: a lifetime learning model. *Developmental Psychology*, *47*, 619–631.

Segal, N. L., & Hershberger, S. L. (1999). Cooperation and competition between twins: Findings from a prisoner's dilemma game. *Evolution and Human Behavior*, *20*, 29–51.

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

Sturgis, P., Read, S., & Allum, N. (2010). Does intelligence foster generalized trust? an empirical test using the UK birth cohort studies. *Intelligence*, *38*, 45–54.

van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*, 796–803.

Verplaetse, 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*, 260–271.

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

Yamagishi, T., Kikuchi, M., & Kosugi, M. (1999). Trust, gullibility, and social intelligence. *Asian Journal of Social Psychology*, *2*, 145–161.

Yang, D., Qi, S. Q., Ding, C., & Song, Y. (2011). An ERP study on the time course of facial trustworthiness appraisal. *Neuroscience Letters*, *496*, 147–151.

Zak, P., & Knack, S. (2001). Trust and growth. *Economic Journal*, *111*, 295–321.