Supplementary Materials

Controlling for Facial Attractiveness and Competence

Perceived trustworthiness naturally co-varies with other traits, such as attractiveness and competence. To eliminate the possibility that attractiveness or competence evaluations spuriously produced our pattern of findings, we recruited three sets of independent raters from Mechanical Turk (each $n = 30$) to rate the 120 face stimuli (60 trustworthy and 60 untrustworthy faces) on either trustworthiness, attractiveness, or competence using a 7-point Likert scale. Interrater agreement was high (intraclass correlation coefficients [ICCs] = .82 to .92). Paired *t*-tests comparing the trustworthy vs. untrustworthy variants of the 60 facial identities confirmed that trustworthy variants were significantly more trustworthy than their untrustworthy counterparts, $t(59) = 10.36$, $p < .00001$, $d = 1.34$. The trustworthy variants were also significantly more attractive, $t(59) = 4.65$, $p < .00001$, $d = 0.60$, which is expected as these traits are typically correlated (Oosterhof & Todorov, 2008). They did not differ in competence, $t(59) = 1.47$, $p = .15$, $d = 0.19$.

To inspect the critical group × trustworthiness interaction effect after including covariates of faces' attractiveness and competence on a trial-by-trial basis, we conducted a meta-analysis of our 6 studies using an aggregated trial-by-trial dataset in a generalized estimating equations (GEE) multi-level regression framework (Riley, Lambert, & Abo-Zaid, 2020). This allowed us to examine whether the training's reduction of facial trustworthiness effects held even when statistically controlling for a given face's attractiveness and competence. Trials were nested within subjects, and subjects were nested within study. The dependent measure in each study was regressed onto trustworthiness (-0.5 = untrustworthy, 0.5 = trustworthy), group (-0.5 = control, 0.5 = trained), and the group × trustworthiness interaction (thus mirroring our primary analyses),

but additionally onto attractiveness (mean-centered), competence (mean-centered), and group × attractiveness and group × competence interactions. The full results are provided in the table below, which reports unstandardized regression coefficients (*B*) and Wald *Z* statistics. There were no significant effects or interactions with attractiveness or competence, and the effect of trustworthiness and, most critically, the group × trustworthiness interaction were still significant and strong when including these covariates in the model.

Effect	***B***	***SE***	**95% CI**		**Wald *Z***	***p***
Trustworthiness	8.6091	1.3415	5.9798	11.2385	6.42	<.0001
Group	-1.0747	1.9963	-4.9873	2.8379	-0.54	.5903
Group × Trustworthiness	-12.2625	2.6831	-17.5212	-7.0038	-4.57	<.0001
Attractiveness	-1.1428	2.4501	-5.9449	3.6593	-0.47	.6409
Competence	3.725	2.6953	-1.5578	9.0077	1.38	.167
Group × Attractiveness	0.3275	4.9002	-9.2766	9.9317	0.07	.9467
Group × Competence	1.6281	5.3907	-8.9374	12.1936	0.3	.7626

Thus, there is no evidence that facial attractiveness or competence confounded the reported pattern of results.

Effects of Participant Race

Our aggregated meta-analytic sample across the 6 studies (see above) consisted of 1,250 total participants. With a large sample in the aggregate, this provided an opportunity to test for possible interactions with participant race. The total sample consisted of 975 White participants (80 of whom identified as Hispanic/Latino), 77 Asian participants (3 identified as Hispanic/Latino), 153 Black participants (20 identified as Hispanic/Latino), and 46 who self-identified as Other (29 identified as Hispanic/Latino). No participants identified as American Indian / Alaska Native or as Native Hawaiian / Other Pacific Islander. We conducted two

analyses, one modeling participant race as White (-0.5) or non-White (0.5) and the other modeling participant race as White (-0.5) or BIPOC (Black, Indigenous, and People of Color). For the latter analysis, BIPOC participants were defined as any participants identifying their race as Black, American Indian / Alaska Native, or Native Hawaiian / Other Pacific Islander (there were no participants in the latter two groups), and any participants identifying their ethnicity as Hispanic/Latino. Thus, for the latter analysis, participants who were not White or BIPOC were excluded. In each case, the dependent measure in each study was regressed onto trustworthiness (-0.5 = untrustworthy, 0.5 = trustworthy), group (-0.5 = control, 0.5 = trained), participant race (-0.5 = White, 0.5 = non-White or BIPOC), and their interactions using the same GEE regression framework as above (trials nested within subjects, which were nested within study). We report unstandardized regression coefficients (*B*) and Wald *Z* statistics.

The full results of the first model defining participant race as either White or non-White are provided in the table below. There was no significant effect of participant race or interactions with participant race.

Effect	***B***	***SE***	**95% CI**		**Wald *Z***	***p***
Trustworthiness	9.6814	1.5924	6.5603	12.8024	6.08	<.0001
Group	0.1714	2.5260	-4.7794	5.1222	0.07	.9459
Group × Trustworthiness	-13.7582	3.1848	-20.0003	-7.5162	-4.32	<.0001
Race	1.5021	2.5260	-3.4487	6.4530	0.59	.5521
Trust × Race	2.4904	3.1848	-3.7517	8.7325	0.78	.4342
Group × Race	4.6219	5.0520	-5.2798	14.5235	0.91	.3603
Trustworthiness x Race x Group	-6.5297	6.3696	-19.0139	5.9544	-1.03	.3053

The full results of the second model defining participant race as either White or BIPOC are provided in the table below. There was no significant effect of participant race or interactions with participant race.

Effect	***B***	***SE***	**95% CI**		**Wald *Z***	***p***
Trustworthiness	10.1315	1.5777	7.0391	13.2238	6.42	<.0001
Group	-1.2715	2.4414	-6.0565	3.5135	-0.52	.6025
Group × Trustworthiness	-11.8848	3.1555	-18.0695	-5.7001	3.77	.0002
Race	0.4395	2.4414	-4.3455	5.2245	0.18	.8571
Trust × Race	1.6897	3.1555	-4.4949	7.8744	0.54	.5923
Group × Race	-1.0038	4.8828	-10.5738	8.5662	-0.21	.8371
Trustworthiness x Race x Group	-1.7628	6.3110	-14.1321	10.6065	-0.28	.7800

Analysis of Training Behaviors

The 10 trustworthy and 10 untrustworthy behaviors used in the training portion of our studies were adapted from a previous study (Lick, Alter, & Freeman, 2017). To confirm that the behaviors conveyed the intended level of trustworthiness, an independent set of raters recruited from Mechanical Turk (n = 30) were asked to rate the trustworthiness of each of the 20 behaviors on a 7-point Likert scale. Interrater agreement was high (ICC = .93), and thus we averaged raters' judgments into a mean trustworthy rating for each behavior. Indeed, the trustworthy behaviors (M = 5.00) were rated as significantly more trustworthy than the untrustworthy behaviors (M = 2.93), $t(18) = 16.22$, $p < 0.0001$, $d = 7.64$. Calculating the distance from the midpoint (4, on our 7-point Likert scale) showed that the trustworthy and untrustworthy behaviors were balanced in extremity and did not significantly differ in distance from the midpoint, $t(18) = 0.53$, $p = 0.60$, $d = 0.24$.

To evaluate whether the behaviors differed on other dimensions that naturally co-vary in real-world trustworthy/untrustworthy behaviors, we recruited four additional sets of independent raters from Mechanical Turk (each $n = 30$) to rate either the arousal/intensity, competence, dominance, and typicality of the behaviors each on a 7-point Likert scale. Interrater agreements were high (ICCs = .74 to .87). The data showed that, relative to untrustworthy behaviors, the trustworthy behaviors were rated as significantly less arousing/intense (Ms = 4.01 vs 4.80; $t(18) = 3.65$, $p = 0.0018$, $d = 1.72$), more typical (Ms = 4.10 vs. 3.52; $t(18) = 3.35$, $p = 0.0036$, $d = 1.58$), less dominant (Ms = 3.81 vs. 4.60; $t(18) = 3.57$, $p = 0.022$, $d = 1.68$), and more competent (Ms = 4.65 vs. 3.65; $t(18) = 3.96$, $p = .0009$, $d = 1.87$). This is unsurprising, as these traits are correlated in real-world behaviors and in perceivers' conceptual knowledge (e.g., Stolier, Hehman, & Freeman, 2020). However, note that trustworthiness had by far the largest effect size ($d = 7.64$). To ensure that these co-varying dimensions did not spuriously produce our pattern of results, we conducted a replication study, described below.

Replication of Study 1B

We conducted a replication of Study 1B to demonstrate that using a balanced set of behaviors in the training would not significantly alter the overall pattern of results in our studies. We chose to replicate Study 1B in particular because, although the trust game used in Studies 1A/1B has been used in several past studies, there may be potential concerns about participants' suspicion or disbelief about the cover story (i.e., that they were presumably playing with real human players). Thus, this replication provided an additional opportunity to add a suspicion probe following the completion of the task. The decision to replicate Study 1B using real faces

rather than Study 1A using computer-generated faces was arbitrary, except for the increased ecological validity of using real faces.

First, to balance the 10 trustworthy vs. 10 untrustworthy behaviors, we generated 8 new behaviors (thus, of the 20 total behaviors, 8 were new and 12 remained identical). We recruited five independent sets of raters from Mechanical Turk (each $n = 30$) to rate the trustworthiness, arousal/intensity, competence, dominance, and typicality of this new set of 20 behaviors each on a 7-point Likert scale. Interrater agreements were high (ICCs = .82 to .92). The trustworthy behaviors were rated significantly more trustworthy than the untrustworthy behaviors (*M*s = 5.03 vs. 2.82; $t(18) = 19.72$, $p < 0.0001$ $d = 9.30$). Critically, however, relative to untrustworthy behaviors, the trustworthy behaviors did not significantly differ on arousal/intensity (*M*s = 4.13 vs. 4.34; $t(18) = 0.79$, $p = 0.44$, $d = 0.37$), typicality (*M*s = 3.71 vs. 3.59; $t(18) = 0.87$, $p = 0.41$, $d = 0.11$), dominance (*M*s = 4.11 vs. 4.26; $t(18) = 0.44$, $p = 0.66$, $d = 0.21$), or competence (*M*s = 4.24 vs. 3.83; $t(18) = 1.73$, $p = 0.10$, $d = 0.82$). Thus, trustworthiness was the only factor that distinguished the two types of behaviors, creating a balanced set.

Using this new balanced set of behaviors, we conducted a replication of Study 1B (the trust game using real faces). Using the same target sample size as the original Study 1B, we recruited 217 total participants from Amazon Mechanical Turk. 11 participants in the control group were excluded (5 for failing any attention check, 6 for having >50% trials register as timeouts during the learning phase) and 7 participants in the trained group were excluded (5 for failing any attention check, 2 for having >50% trials register as timeouts during the learning phase). There were 101 total participants in the final control group (age: M=35.8 years, SD=8.9 years; 61 male; race: 71 White, 25 Black, 4 Asian, 1 Native American; Hispanic/Latino ethnicity: 19) and 98 total participants in the final trained group (age: M=37.7 years, SD=11.7

years; 55 male; race: 64 White, 31 Black, 3 Asian; Hispanic/Latino ethnicity: 31). The procedures were identical to the original Study 1B with two exceptions: 1) we used the new balanced trustworthy and untrustworthy behaviors as part of the training, and 2) following the completion of the study and prior to debriefing, we probed participants using a free-response question: "Did you notice anything unusual or inconsistent about the experiment?"

Trust game payment was analyzed using a 2 (study: original study vs. replication) x 2 (group: control vs. trained) x 2 (facial trustworthiness: trustworthy vs. untrustworthy) mixed-model ANOVA. There was a main effect of facial trustworthiness, $F(1,397) = 64.78$, $p < .0001$, $\eta_p^2 = 0.14$, with more money paid to targets with trustworthy than untrustworthy faces. There was no main effect of group, $F(1,397) = 0.001$, $p = .981$, $\eta_p^2 < 0.001$. More importantly, there was a significant trustworthiness × group interaction, $F(1,397) = 44.63$, $p < .0001$, $\eta_p^2 = 0.10$. Crucially, there was no main effect of study or any interactions with study: study, $F(1,397) = 0.05$, $p = .83$, $\eta_p^2 < 0.001$; study × trustworthiness, $F(1,397) = 0.39$, $p = .53$, $\eta_p^2 = 0.001$; study × group, $F(1,397) = 0.02$, $p = .89$, $\eta_p^2 < 0.001$; study × group × trustworthiness; $F(1,397) = 0.42$, $p = .52$, $\eta_p^2 = 0.001$. Replicating the trustworthiness × group interaction, and the lack of any interactions with study, show that the pattern of results obtained using either the original behaviors or the new balanced behaviors were statistically indistinguishable. Thus, having balanced or unbalanced behaviors did not have an appreciable effect on the results.

In the suspicion probe, a plurality of responses indicated no evidence of any systematic suspicion. Only 2/200 participants mentioned anything potentially concerning. For instance, one mentioned that "the faces loaded slowly, which seemed like a waste of time." The subject was referring to our random latency period in between trials, simulating the time it would take to match with the next player. The other subject (in the control condition) mentioned that the faces

that they encountered seemed to be trustworthy or untrustworthy. Even when removing these 2 participants, all the reported results still held. Thus, there was no evidence that generally participants were suspicious or did not believe that they were playing with other humans.

References

Lick, D. J., Alter, A. L., & Freeman, J. B. (2018). Superior pattern detectors efficiently learn, activate, apply, and update social stereotypes. *Journal of Experimental Psychology: General*, *147*(2), 209.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087-11092.

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *Bmj*, 340, c221.

Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature human behaviour*, *4*(4), 361-371.