**Peer Review and Communication History**

**MS Title**: Are Natural Faces Merely Labelled as Artificial Trusted Less?

**Author Names**: Baptist Liefooghe, Manuel Oliveira, Luca M. Leisten, Eline Hoogers, Henk Aarts1, Ruud Hortensius

**Submitted:** June 10, 2022

**Editor First Decision**: Revise & Resubmit

Aug 17, 2022

Dear Oliveira,

I have now received XX reviews of your manuscript, “Faces Merely Labelled as Artificial are Trusted Less”, from researchers with special expertise in XXX. I also independently read the manuscript before consulting these reviews. The reviewers had mixed reactions to your manuscript. I agree that your manuscript has important strengths and also that an additional experiment is needed to fully address the concerns. If you are willing to conduct this study, I therefore encourage you to submit a revised version for further consideration at Collabra: Psychology.

The reviewers did an outstanding job in their reviews. I will highlight issues I think are particularly salient here. In your resubmission, please include a document with a point-by-point response to both the points I list here and the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.

1. One of my major concerns, also noted by Lisa DeBruine (who signed their review), was with the way that the face stimuli were selected for the three experiments. These were selected such that they fell into high/low attractiveness/trustworthiness quadrants using ratings provided by the Chicago Face Database. According to the manuscript, this approach was intended to control whether the effect of label depends on the degree of attractiveness and trustworthiness associated with the face. However, this creates two problems. First, because trustworthiness and attractiveness are correlated, the high-low combinations may be unusual to participants and might elicit responses unrepresentative of faces in general. Second, as DeBruine carefully details, these categories are confounded with gender and ethnicity. Although this confound would be more serious if the main manipulation – label type – were confounded with gender and ethnicity, I still think these selection issues are quite serious. They undermine the conclusions drawn regarding the trustworthiness and attractiveness manipulations, and they also reinforce the harmful notion that white faces are more attractive than black faces. Although one might argue, to this latter point, that this pattern in the selected faces simply captures the biases of the raters (if that is, in fact, the case), this is a methodological choice that requires transparency and careful explanation. Due to the interpretive challenges introduced by this confound, I think this paper can only achieve its intended goals with the addition of a fourth experiment that resolves these selection issues.
2. Both DeBruine and Reviewer 2 express concerns about the use of ANOVA for data analysis. DeBruine provides specific advice about how to do these analyses using cross-classified mixed effects models. This approach would bring your analyses more in line with the types of general conclusions you draw in the manuscript.
3. Reviewer 2 and I both wondered if the within-subjects design might allow participants to guess the purpose of the study and create contrast or demand effects. These possibilities should be addressed.
4. On a conceptual level, I wondered: When asked to judge the trustworthiness of an artificially generated face, what are participants actually judging? It seems possible that they are judging an imagined computer that is linked to that face (since it can’t be a human). So, might the findings reported here reflect greater distrust of computers compared to people? This seems like a slightly different explanation than those provided in the manuscript, and one that could be worth considering.

In summary, I think this is a promising manuscript and, I hope you will consider running an additional experiment to address the concerns. Please see the instructions below for submitting your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This may be the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.

We hope you can submit your revision within the next three months. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Alexa Tullett

# Reviewer 1

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  | ✔ |  |  |

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

I’m going to start with the TL;DR conclusion: I think this paper uses generally rigorous methods to address an interesting question. I have some advice about the analysis that I think will strengthen the paper. See DeBruine & Barr (2021 AMPPS) for explanation of the benefits of cross-classified mixed effects models for data like yours. This blig post also has concrete examples (<https://debruine.github.io/post/aggregating/>). In brief, when you aggregate across observations you ignore the sampled nature of your face stimuli, so can only make conclusions about this exact image set, rather than generalise to other image sets.

My main concern about the paper that can’t be addressed by additional analyses is the selection of faces as high and low attractiveness and trustworthiness. These two perceptions tend to be positively correlated, so your distribution of faces is unnatural and the high/low combos may be unusual in some way. More importantly, they are confounded for gender and ethnicity, e.g., Black faces were twice as likely to be in the low attractiveness groups (8) compared to the high attractiveness groups (4), while white faces were five times more likely to be in the high attractiveness groups (5) than the low attractiveness groups (1). This should, at the very least, warrant discussion, and at the most, prompt a further study looking at faces with more balanced demographics (or more homogeneous faces across the natural ranges of trustworthiness and attractiveness).

Signed: Lisa DeBruine
(please feel free to contact me if you have any questions about my review)

## Introduction

P. 3 “Trustworthiness judgments on the basis of faces follow from our expertise in inferring traits from facial features, …“

It’s important to note for context that there is little to no evidence that trustworthiness judgements, as consistent as they are, are related to actual trustworthy behaviour.

See work by Jones et al (2001, Nature Human Behaviour) for evidence for agreement and cross-cultural consistency of trustworthiness judgements, and reviews by Todorov on the relationship to behaviour.

## Experiment 1

### Methods

P.7 “To obtain a power of .80 for a medium-sized effect (d= 0.5) in a within-subjects design, a minimum sample of 32 participants was needed.”

What is the justification for expecting the effect size to be 0.5? It might be more informative here then to report what was the minimum effect size you had 80% power to detect with this sample size.

P. 8 - “we selected faces with the most extreme scores on both dimensions such that we obtained four categories of 6 faces each”

Could you be more clear about the algorithm used for this? The preregistration for Experiment 3 states: “cut-off values of lower than 3.5 corresponding to low, and higher than 3.5 corresponding to high”, but here you say most extreme scores. Since you’re selecting for extremes on 2 dimensions, did you do some sort of cluster analysis? (e.g., how would you choose between a face that’s 2.1 trust and 5.4 att, vs 2.2 trust and 5.5 att?) I know I’m being picky here, but it’s important for potential replication efforts to be clear on your stimulus selection criteria.

P. 8 - You used 24 faces that were Black or White and male or female, divided into 4 groups of high/low attractiveness trustworthiness. Looking at the supplemental materials, it seems that these groups were not balanced for gender or ethnicity (and the set also included Asian and Latina faces):

| Att | Trust | BF | BM | WF | WM | LF | LM | AF | AM |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| High | High | 1 | 1 | 3 |  | 1 |  |  |  |
| High | Low | 1 | 1 | 1 | 1 | 2 |  |  |  |
| Low | High | 2 | 2 |  |  | 1 |  |  | 1 |
| Low | Low | 3 | 1 | 1 |  | 1 |  |  |  |

Thus introduced some unacceptable confounds, such as Black faces being twice as likely to be in the low attractiveness groups (8) compared to the high attractiveness groups (4), while white faces were five times more likely to be in the high attractiveness groups (5) than the low attractiveness groups (1).

The supplemental materials were a little confusing at first because the last line is drawn in the wrong place.

What was the ethnic make-up of the observers in this study? If not specifically collected, what is the likely make-up, given the population recruited from? This is important, given the potential for outgrip effects in face perception.

### Results

* Please indicate explicitly in the text that data were aggregated by observer (unless you change the analysis to a more appropriate mixed effects model). Although you didn’t explicitly do this in the code, apex does this for you (e.g., try explicitly calculating the mean for the 4 groups for each person — the ANOVA results will be identical)
* An interaction plot of the trustworthiness ratings for the 4 groups would be useful for interpretations here

I very strongly recommend using mixed effects model to analyse these data instead of an ANOVA. This will account for the fact that the face stimuli are sampled, in the same way that observers are sampled from a larger population that you wish to generalise to. This is additionally important because of the relatively small number of faces and imbalance of face characteristics between the groups. Otherwise, you can only make conclusions that generalise to the population of observers, but not ones that generalise beyond the sample of 24 stimulus faces you used.

However, this will require you to add Face ID labels to each trial. It is currently not possible to determine which face was shown on each trial, or I would have run this analysis for you to demo. See DeBruine & Barr (2021, AMPPS) for examples of a cross-classified mixed effects model similar to your study.

### Code

* The dependency files setup.R and helpers.R are not included in the shared code.
* I could figure out most of the package dependencies, but couldn’t run the function interactionMeans() — I found compute\_alpha() in the Experiment 3 helpers.R file
* I could run the rest of the code, but did not check whether the values match the manuscript, as I will be recommending a different analysis, but it really helps me do reproducibility code check if you make it very clear how the analyses in the code correspond to the numbers in the results section. A great way to do this is to use sprintf() or glue() or inline R to create the text of the results section with inserted values from code.
* The data processing code is very hard to understand and needs comments (e.g., I can’t figure out why you created and add the column man\_check, as it’s identical to the existing Check column)
* FYI: the code from lines 55-79 could be condensed into
test=which( experiment\_l$stimulus %in% c( ‘A\_3\_1’, …rest of the values…, ‘D\_2\_2’) )

## Experiment 2

### Methods

* Same methodological critiques as exp 1

### Results/Code

* Same advice as Exp 1 about mixed models
* Same problem with missing helper files
* I could run everything apart from the code that used interactionMeans()
* I am not really sure what that cronbach’s alpha is doing. You seem to be averaging the ratings over the six faces in each unique value of participant\_nr, story, attractiveness, trustworthiness, likert\_scale, and label, and then calculating the Alph for the 8 means per participant grouped by likert\_scale (judgement) and story (i.e., each combo of att/trust/label). This could use more clarity with comments in the code, as it’s not discussed in the manuscript at all. (Also relevant to Exp 1)

## Experiment 3

P. 14 “Experiment 3 aims to replicate the findings of Experiment 2 with sufficient statistical power to detect a potentially smaller (i.e., d = 0.4) moderation of the bias towards computer-generated faces by background story.“

With 60 subjects, you have 80% power to detect d = 0.36 in a paired design, so Experiments 1 and 2 were already sufficiently powered for this.

### Methods

I commend you for including a preregistration (<https://osf.io/w4bca>). It is very clear and detailed (seriously, I’m impressed!). This isn’t always the case, in my experience, so please include a clearer summary of what was preregistered and what deviations there were (e.g., the adjustment to the minimum time allowed to be spent on the story screen mentioned in the analysis script). Quite often, I see papers claim to be preregistered in a way that insinuates everything was specified, and after looking at the registration document find out they only roughly specified half the hypotheses and said they’d be analysed “with anova” without further details. Make sure a reader knows how detailed your pre-reg is in methods and analyses (and then they can get more details at the link if needed).

P. 16 “The previously used 24 pictures (6 per face category; see Experiment 1) were complemented with 18 faces from the same database per category, adding up to a total of 96 faces.“

Please include the face image IDs for these extra faces, as well as their values for the category ratings.

### Results

I’d also recommend adding a cross-classified mixed effects model for this analysis. Instead of calculating d’ for the memory data, you can use a binomial glmer. This will account for the randomising of face stimulus identities between subjects, as well as for stimulus sampling.   As the analyses here are pre-registered, you should definitely report the original planned analysis and compare them with the mixed effects analyses as robustness checks.

### Code

Experiment 3 needs a README to tell the user what order to run the scripts in. I assumed Anonymization, Cleaning, then Analysis. Obviously, Data Anonymization won’t have the raw data shared, so please also explain this explicitly in the README.

The Data Cleaning script has very good comments/text and is much easier to follow than the scripts for Exp 1 and 2.

* Line 44 reads the data from the directory “data”, but this is actually called “anonymized raw data” in your shared materials.
* This script ran fine after fixing this
* I skimmed the code for sense and didn’t see anything obviously wrong, but didn’t do line-by-line code check

The Data Analysis code ran fine without changes 😃

* The long narrative comments (e.g., lines 38-56) could be moved out of the code chunks into the markdown section of the scripts. This usually makes the HTML output easier to read. This is just stylistic advice, but I tend to put scientific narrative in the markdown and use code comments only for specifically code-relevant notes.
* The alpha code produces a LOT of messages: “In smc, smcs < 0 were set to .0”; it might be useful to hide those and add a comment explaining why.
* In the Data Processing section, it’s not 100% clear why you drop\_na() and loose about 6000 observations, but these are all observations with NA in the label column. It would be useful if this section had a little more narrative about the processing.
* I didn’t have capacity to do line-by-line code check

## General

Figure 1 implies the Likert ratings were a sliding scale, but the data show values as integers. Also, the figure doesn’t match the methods text on Page 9: “Below each face, two 7-point Likert scales were presented. One for attractiveness and one for trustworthiness. For both scales the right side was labeled with “absolutely not” (1) and the right side with “extremely” (7). “

I’m a little confused by Figure C. The y-axis is labelled observations, which I would normally assume are integers, but the histograms start at 0.1 and go as high as 0.6, so I assume this is either some type of density plot or an artifact of plotting the horizontal box plots on the same plot. The x-axis is perceived trustworthiness, with labels 2, 4 and 6, but the histogram bars seem to be 0.5 in width, so are these plotting mean rating scores or individual observation ratings?

I spent 20 minutes installing new packages (some dependencies had to be installed from source), so make sure you’re only calling what you need from the setup.R file. Thanks for including the install code for non-CRAN packages 😃

FYI: You can resolve functions conflicts by loading the package with default functions last (e.g., load tidyverse last to use dplyr functions as defaults). This avoids functions in the global environment. (But is more a stylistic choice and not a problem.)

# Reviewer 2

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  | ✔ |  |  |

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The manuscript *“Faces Merely Labelled as Artificial are Trusted Less”* investigates whether the relationship between AI and people is in the mind of the beholder. In three experiments, the authors showed that participants perceived natural faces that were merely labelled as artificial as less trustworthy than natural faces labelled as natural. This is an interesting and relevant topic considering the increasingly pervasive presence of computer-generated faces in our everyday life. These results open a new question: why do people trust computer-generated faces less than natural faces? Previous research had hypothesised that people are less expert in inferring traits from artificial faces because we do not see them quite as often as natural faces, but this study showed that it is possible to modulate people’s judgment also when facial features are natural. A (pre-registered) working hypothesis made by the authors was that artificial faces would have been considered as part of an outgroup. This was measured using a memory task, but the null hypothesis could not be rejected.
I generally liked the manuscript which reads very nicely and it is quite clear in every section. I very much appreciate the pre-registration for study 3 and the fact that all scripts and data are shared already. The manuscript is well-written, relevant and opens new interesting questions about the relationship between AI and humans.
I have a few major clarifications and some minor suggestions to hopefully improve the readability of the manuscript.

**MAJOR POINTS**
Major 1:
As I said, I very much enjoyed reading the manuscript, but I feel the results would have been easier to grasp using more appropriate figures. As it is, there is only one figure and the subpanels are not even properly referred to, unless I missed it. It would be very useful to see the results of the ANOVAs, and possibly to show individual participants to have an idea of the sample distributions. The description of the results of the ANOVAs is great, but some readers (like me) would prefer to see them also in a plot.

Major 2:
I was wondering whether the type of story that you used could have somehow revealed the aim of the study. Did people understand that you wanted to find a difference between artificially generated faces and natural faces for the two stimulus characteristics? I am not saying that this is the case, but I wanted to get your opinion on this. It is true that you don’t see the same effect on attractiveness for experiment 1 and 2, and this could be a good sign that your experimental manipulation was specific to trustworthiness. However, attractiveness could be more difficult to manipulate without manipulating the perceptual aspect of the stimuli. At the same time, you do see an effect on attractiveness in experiment 3. I would just like to hear if you think the possibility of this type of bias can be ruled out, and eventually add a sentence or two in the discussions.

Major 3:
I was initially suspicious about your choice of using the ANOVA to analyse your data, because you are essentially only treating the participants as a random effect and the stimuli as a fixed effect which could inflate Type I error (see Judd et al., 2012, J Personality and Social Psychology). However, due to the nature of your stimuli (i.e., they are always the same faces) and the fact that you counterbalanced the labels (computer-generated and natural) between participants, this might not be a problem, as any effect due to the stimuli should cancel out. In any case, using the ANOVA makes the analysis a bit complicated (e.g., in experiment 2 and 3 you have a 4 factors design) and you cannot control directly for covariates (e.g., gender, age, attractiveness judgment). Wouldn’t have been just simpler and more appropriate to use mixed-effects models here? Can you otherwise rule out that the effects that you see are not due to differences in gender, age, or attractiveness judgments?

Major 4:
I was also suspicious about the choice of using ANOVA because your dependent variable is not continuous but it is ordinal. It is probably common to use ANOVA in this situations, but this does not justify using it. Have you checked that the assumptions of the ANOVA are met?

**MINOR POINTS**
Minor 1:
Panel D seems a nice figure, but I don’t think it is referred to in the manuscript.

Minor 2:
“Attractiveness judgements are mainly based on a global affective response that requires minimal inferential activity (e.g., Zajonc, 1980) and offers a benchmark for more sophisticated judgments such as of trustworthiness (see also Willis & Todorov, 2006).”
What do you mean with “offers a benchmark” here? And why did you choose attractiveness in this experiment?

Minor 3:
Related to Minor 2. Why you used the attractiveness associated to each face instead the attractiveness estimated by the participants?

Minor 4:
“In each category, faces were randomly divided in two sets of 3 faces. These sets were either labelled as being natural or computer-generated faces. This labelling was counterbalanced over participants.”
Could you please clarify whether there were always the same 3 faces in each set?

Minor 5:
“Participants were either first presented with the set of faces labelled as natural or with the set of faces labeled as computer-generated. Each set consisted of 12 faces (3 faces per category). Within each set faces were presented in a random order, one at a time in the middle of the screen.”
Could you please clarify whether also the category sets were randomly presented?

Minor 6:
“Below each face, two 7-point Likert scales were presented. One for attractiveness
and one for trustworthiness. For both scales the right side was labeled with “absolutely not” (1) and the right side with “extremely” (7).”
From Figure 1A, it seems that the scale was continuous, not a Liker scale, and the labels were “Not at all” to “Extremely”. Which ones are correct?

Minor 8:
“Participants were also informed about the nature of the upcoming set of faces (i.e., natural or computer-generated).”
When was this presented? Was it before each face? Before each set? Always present? Please clarify.

Minor 9:
Regarding Figure 1B, could you show the percentage rather than the absolute values? I think it would facilitate the comparisons between experiments. How do you explain the fact that many people (especially in experiment 3) did not believe that the faces were artificial, but you can still see an effect on trustworthiness? This might be related with Major 2, meaning that the aim of the study was evident to participants.

**Author Response**
Dec 8, 2022

Dear Professor Alexa Tullet,

With great pleasure, we submit for your consideration our revised article entitled “Faces Merely Labelled as Artificial are Trusted Less”. Below, we briefly summarize our responses to the main points raised during peer-review.

We would like to thank the reviewers and you for their constructive and supportive comments. We have revised the manuscript according to their excellent suggestions and detail our responses in the comments below. In sum, our major revisions include:

* The inclusion of two additional preregistered experiments (Experiments 4 and 5) that address the issue of stimuli selection (Experiment 5, point 1), contrast or demand effects (Experiment 5, point 3), and the conceptual point (Experiment 4, point 4).
* The use of cross-classified mixed effects models throughout all the five experiments (point 2)

Again, we want to thank the reviewers for providing us with valuable feedback that allowed us to strengthen our paper. We thank you for the chance to resubmit our work to *Collabra: Psychology* and look forward to hearing back from you and expert peer reviewers regarding the suitability of this work for publication in *Collabra: Psychology*.

With best regards,

And on behalf of all the co-authors,

Authors

**Editor’s comments and authors’ responses.**

**Comment 1.** *One of my major concerns, also noted by Lisa DeBruine (who signed their review), was with the way that the face stimuli were selected for the three experiments. These were selected such that they fell into high/low attractiveness/trustworthiness quadrants using ratings provided by the Chicago Face Database. According to the manuscript, this approach was intended to control whether the effect of label depends on the degree of attractiveness and trustworthiness associated with the face. However, this creates two problems. First, because trustworthiness and attractiveness are correlated, the high-low combinations may be unusual to participants and might elicit responses unrepresentative of faces in general. Second, as DeBruine carefully details, these categories are confounded with gender and ethnicity. Although this confound would be more serious if the main manipulation – label type – were confounded with gender and ethnicity, I still think these selection issues are quite serious. They undermine the conclusions drawn regarding the trustworthiness and attractiveness manipulations, and they also reinforce the harmful notion that white faces are more attractive than black faces. Although one might argue, to this latter point, that this pattern in the selected faces simply captures the biases of the raters (if that is, in fact, the case), this is a methodological choice that requires transparency and careful explanation. Due to the interpretive challenges introduced by this confound, I think this paper can only achieve its intended goals with the addition of a fourth experiment that resolves these selection issues.*

**Response:**

Thank you for this thoughtful comment. We agree that the selection of the faces was not optimal in Experiments 1 and 2. Specific combinations of attractiveness and trustworthiness may have undermined specific interactions with our manipulation of label (“computer-generated” vs. “natural”). However, the same faces were used in both label conditions. Hence, the main effect of label seems unlikely to be affected by the specific faces used. In Experiment 3, a more elaborate randomization procedure was used. 96 faces were selected from the Chicago Face Database, 24 per category. For each participant, 6 faces were randomly selected for each of the four categories. In contrast to Experiments 1-2 the effect of specific combinations of attractiveness and trustworthiness was thus mitigated. In addition, we reanalyzed the data by using linear mixed models in which random intercepts were estimated for the faces. Hence, the variance of the faces was now accounted for by our analyses. However, the same results were obtained. Finally, in Experiment 4 we used a completely different set of (talking) faces, which now were neutral with respect to trustworthiness. However, we still replicated the same effects. To conclude, we believe that we have taken the concern of the reviewers into consideration by analyzing our data differently and adding new data.

We have also added this issue in the GD of the manuscript (pp. 35-36):

“A common limitation in the face perception literature is that stimulus materials are often relatively restricted with respect to the ethnicity of the faces used (Cook & Over, 2021). In the present study, our stimulus sets encompassed several ethnicities (Experiments 1-3 and 5), which is more in line with the contemporary multi-ethnic reality (e.g., Hong & Cheon, 2017). However, for the exception of Experiment 5, our experiments were not fine-tuned to counterbalance the proportions of different face ethnicities in the stimulus sets. The absence of counterbalancing introduced some biases in the stimulus sets where stimuli were selected on the basis of attractiveness and trustworthiness ratings (Experiments 1-3). Such ratings may reflect stereotype-driven biases (Lewis, 2011; Schmid et al., 2022; Sutherland et al., 2015) that result into a disproportionate amount of faces of a particular ethnicity to fall into a specific level of the judgment dimension (e.g., more White faces than Black faces in high attractiveness categories; but see Lewis, 2011). Nevertheless, this limitation is unlikely to have been the main source of the label effect for several reasons. First, we replicated the label effect under different conditions of stimulus variability (unlike in Experiments 1, 2, and 5, each participant in Experiment 3 was exposed to a set composed of different faces). Second, the effect also emerged in Experiment 4 with a different stimulus conveying a richer set of social cues, despite of its lower diversity (i.e., French-speaking natives of European descent) and imbalanced face sex (although consistently so cross label conditions). Moreover, the stimuli in Experiment 4 were manipulated to be more homogeneous in perceived trustworthiness, thus preventing the disproportionate allocation of any ethnicity to specific levels of a judgment. Finally, although it is tempting to conclude that the introduction of the optimal counterbalancing of social cues in Experiment 5 may have contributed to reduce or eliminate the label effect, such an explanation would remain harder to conciliate with the effect found with a sample that was more homogeneous in terms of physical appearance and perceived trustworthiness in Experiment 4. Instead, we believe that a contrast effect and/or lower believability in the nature of the stimuli in the computer-generated condition, are more likely to have been the factors resulting in the absence of a label effect in Experiment 5.”

**Comment 2.** *Both DeBruine and Reviewer 2 express concerns about the use of ANOVA for data analysis. DeBruine provides specific advice about how to do these analyses using cross-classified mixed effects models. This approach would bring your analyses more in line with the types of general conclusions you draw in the manuscript.*

**Response:**

Thank you for this excellent recommendation. We re-analyzed all data using linear mixed effect models and have updated all result sections accordingly. The outcome of these analyses converged with our previous analyses and conclusions.

**Comment 3.** *Reviewer 2 and I both wondered if the within-subjects design might allow participants to guess the purpose of the study and create contrast or demand effects. These possibilities should be addressed.*

**Response:**

We added a fifth experiment in which this hypothesis was tested. The results indicate that this is indeed the case. However, alternative conclusions are possible. We believe this does not undermine our study but is an important issue in research on the comparison between AI and humans in general. We offer a discussion of this issue in the Discussion of Experiment 5 (p. 31) and the GD (p. 30 and 32):

Experiment 5 Discussion:

“In contrast to the previous experiments, we did not observe a label effect when faces were presented to two groups of participants, which only rated one type of face. This finding thus suggests that the label effect is possibly only present in contexts which require the judgment, or the presence of both faces labelled as natural and artificial.”

General Discussion

p.34

“Finally, Experiment 5 suggests that this impression formation depends on the presence of different category of faces, which may indicate that some comparison or contrast between these categories is needed for the label effect to occur.”

p.36

“Finally, although it is tempting to conclude that the introduction of the optimal counterbalancing of social cues in Experiment 5 may have contributed to reduce or eliminate the label effect, such an explanation would remain harder to conciliate with the effect found with a sample that was more homogeneous in terms of physical appearance and perceived trustworthiness in Experiment 4. Instead, we believe that a contrast effect and/or lower believability in the nature of the stimuli in the computer-generated condition, are more likely to have been the factors resulting in the absence of a label effect in Experiment 5.”

**Comment 4.** *On a conceptual level, I wondered: When asked to judge the trustworthiness of an artificially generated face, what are participants actually judging? It seems possible that they are judging an imagined computer that is linked to that face (since it can’t be a human). So, might the findings reported here reflect greater distrust of computers compared to people? This seems like a slightly different explanation than those provided in the manuscript, and one that could be worth considering.*

**Response:**

We added a fourth experiment to offer first insights about the inferences that participants made when judging the trustworthiness of the faces by using talking faces. Each face said aloud the following message: “It’s ten minutes to two o’clock.” The faces were presented together with a digital clock, which could indicate the same or different time. We assumed that the communication of inaccurate information might affect perceptions of trustworthiness (e.g., by sounding deceitful, or as unreliable). On the one hand, participants may only rate the faces proper. On the other hand, participants may try to make some inferences or hypotheses about the person to which the faces belong and use these to rate the trustworthiness. By letting the faces making correct or incorrect assertions, we wanted to explore whether the label effect interacted with the correctness of the message communicated by the faces. If the label effect is mediated by some inferences about the person behind the face, then the label effect may interact with the correctness of the message. Because this message is part of the behavior of that person. Our results indicated a label effect and an effect of the correctness of the message but no interaction. This finding suggests that trustworthiness ratings are based on the facial appearance alone. We discuss these findings in Experiment 4 (p. 27-28) and in the GD (p. 34):

Experiment 4 Discussion

“Experiment 4 replicates the results of the previous experiments. Faces merely labelled as being artificial are trusted less. Faces indicating the incorrect time were also considered to be less trustworthy compared to faces who indicated the correct time. However, both effects did not interact. The label effect does not seem to be modulated by the ‘behavior’ of the face. Such finding suggests that the label is effect is based on the judgment of the face rather than the judgment of the whole agent . Finally, in Experiment 4 different words were assigned to the two label categories. Nevertheless, we still replicated the label effect. This indicates that the label effect is not limited to the specific pairs of words used in Experiments 1-3.”

General Discussion

“The results of Experiments 1-3 are difficult to reconcile with our initial hypothesis that the faces labelled as being artificial are represented as an outgroup, which results in these faces to be rated as less trustworthy (Balas & Pacella, 2017). At the same time, Experiment 4 further confirmed that the label effect is originated at the early stages of impression formation without the ‘behavior’ of the face coming at play.”

**Reviewer 1**

**General comment:**

*I’m going to start with the TL;DR conclusion: I think this paper uses generally rigorous methods to address an interesting question. I have some advice about the analysis that I think will strengthen the paper. See DeBruine & Barr (2021 AMPPS) for explanation of the benefits of cross-classified mixed effects models for data like yours. This blig post also has concrete examples (*[*https://debruine.github.io/post/aggregating/*](https://debruine.github.io/post/aggregating/)*). In brief, when you aggregate across observations you ignore the sampled nature of your face stimuli, so can only make conclusions about this exact image set, rather than generalise to other image sets.*

*My main concern about the paper that can’t be addressed by additional analyses is the selection of faces as high and low attractiveness and trustworthiness. These two perceptions tend to be positively correlated, so your distribution of faces is unnatural and the high/low combos may be unusual in some way. More importantly, they are confounded for gender and ethnicity, e.g., Black faces were twice as likely to be in the low attractiveness groups (8) compared to the high attractiveness groups (4), while white faces were five times more likely to be in the high attractiveness groups (5) than the low attractiveness groups (1). This should, at the very least, warrant discussion, and at the most, prompt a further study looking at faces with more balanced demographics (or more homogeneous faces across the natural ranges of trustworthiness and attractiveness).*

**Response:**

We thank the reviewer for the constructive comments and we hope that we succeeded in reworking our manuscripts in line with the different suggestions provided. In short, we added two experiments and re-analyzed all data in line with comments provided by the reviewer.

**Comment 1.** *Introduction*

*P. 3 “Trustworthiness judgments on the basis of faces follow from our expertise in inferring traits from facial features, …“*

*It’s important to note for context that there is little to no evidence that trustworthiness judgements, as consistent as they are, are related to actual trustworthy behaviour.*

*See work by Jones et al (2001, Nature Human Behaviour) for evidence for agreement and cross-cultural consistency of trustworthiness judgements, and reviews by Todorov on the relationship to behaviour.*

**Response:** We thank the Reviewer for this suggestion. We now replaced that sentence with a more nuanced summary of what we meant to say in the Introduction section (p. 3):

“People are generally able to rapidly form impressions about someone’s trustworthiness based on their facial appearance alone (B. C. Jones et al., 2021; Oosterhof & Todorov, 2008; Todorov et al., 2009), regardless of the accuracy of such subjective impressions in actually predicting behaviour (for a review see Todorov et al., 2015, pp. 531–535). These appearance-driven judgments are influenced by the degree to which we are exposed to particular distributions of facial features in our environment (e.g., Dotsch et al., 2016; Ng & Lindsay, 1994).”

**Comment 2.**

*Experiment 1*

*Methods*

*P.7 “To obtain a power of .80 for a medium-sized effect (d= 0.5) in a within-subjects design, a minimum sample of 32 participants was needed.”*

*What is the justification for expecting the effect size to be 0.5? It might be more informative here then to report what was the minimum effect size you had 80% power to detect with this sample size.*

**Response:**

Thanks for your suggestion. Indeed a sensitivity aanalysis does seem more appropriate to report on second thought. We rewrote the power considerations for Experiment 1 as (p.6):

“This sample size (N = 60) allows to detect an effect as small as d = 0.465, with 80% power and alpha = .05, for a within-subjects design (with 24 target stimuli nested within condition). The sample size is thus sufficiently large for our research purpose.”

**Comment 3.** *P. 8 - “we selected faces with the most extreme scores on both dimensions such that we obtained four categories of 6 faces each”*

*Could you be more clear about the algorithm used for this? The preregistration for Experiment 3 states: “cut-off values of lower than 3.5 corresponding to low, and higher than 3.5 corresponding to high”, but here you say most extreme scores. Since you’re selecting for extremes on 2 dimensions, did you do some sort of cluster analysis? (e.g., how would you choose between a face that’s 2.1 trust and 5.4 att, vs 2.2 trust and 5.5 att?) I know I’m being picky here, but it’s important for potential replication efforts to be clear on your stimulus selection criteria.*

**Response:**

The Reviewer is correct in pointing out this weakness. Experiments 1 and 2 were conducted during the bachelor thesis of the fourth author. The heuristic used was rather lenient and this motivated us to conduct a replication in Experiment 3. At the same time, we clearly shared information about the CFD stimuli used in Experiments 1 and 2 (see Face stimuli info at the OSF repository: https://osf.io/32tkj). Hence, we believe that a replication of these experiments is possible.

**Comment 4.** *P. 8 - You used 24 faces that were Black or White and male or female, divided into 4 groups of high/low attractiveness trustworthiness. Looking at the supplemental materials, it seems that these groups were not balanced for gender or ethnicity (and the set also included Asian and Latina faces):*

| *Att* | *Trust* | *BF* | *BM* | *WF* | *WM* | *LF* | *LM* | *AF* | *AM* |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *High* | *High* | *1* | *1* | *3* |  | *1* |  |  |  |
| *High* | *Low* | *1* | *1* | *1* | *1* | *2* |  |  |  |
| *Low* | *High* | *2* | *2* |  |  | *1* |  |  | *1* |
| *Low* | *Low* | *3* | *1* | *1* |  | *1* |  |  |  |

*Thus introduced some unacceptable confounds, such as Black faces being twice as likely to be in the low attractiveness groups (8) compared to the high attractiveness groups (4), while white faces were five times more likely to be in the high attractiveness groups (5) than the low attractiveness groups (1).*

**Response:**

Thank you for identifying this issue and investing time to clearly expose it in a table. This limitation is now acknowledged in the General Discussion (p. 31-32, see below). This issue was also mitigated in the two additional experiments. In Experiment 5 we counterbalanced the ethnicity and sex of the stimuli. In Experiment 4, we used different stimuli which were less diverse in ethnicity, but the stimuli were filtered to include those rated in the middle range of trustworthiness (average of face and voice trustworthiness) according to the original ratings of their respective database.

General Discussion p.35-36

“A common limitation in the face perception literature is that stimulus materials are often relatively restricted with respect to the ethnicity of the faces used (Cook & Over, 2021). In the present study, our stimulus sets encompassed several ethnicities (Experiments 1-3 and 5), which is more in line with the contemporary multi-ethnic reality (e.g., Hong & Cheon, 2017). However, for the exception of Experiment 5, our experiments were not fine-tuned to counterbalance the proportions of different face ethnicities in the stimulus sets. The absence of counterbalancing introduced some biases in the stimulus sets where stimuli were selected on the basis of attractiveness and trustworthiness ratings (Experiments 1-3). Such ratings may reflect stereotype-driven biases (Lewis, 2011; Schmid et al., 2022; Sutherland et al., 2015) that result into a disproportionate amount of faces of a particular ethnicity to fall into a specific level of the judgment dimension (e.g., more White faces than Black faces in high attractiveness categories; but see Lewis, 2011).”

**Comment 5.** *The supplemental materials were a little confusing at first because the last line is drawn in the wrong place.*

**Response:**

Thank you for alerting us to this formatting error. We fixed the formatting and updated the pdf file in OSF (CFD\_face\_stimuli\_reference\_codes\_and\_ratings.pdf).

**Comment 6.**

*What was the ethnic make-up of the observers in this study? If not specifically collected, what is the likely make-up, given the population recruited from? This is important, given the potential for outgrip effects in face perception.*

**Response:**

Thank you for the suggestion. Although we did not initially collect data about self-reported ethnicity in our experiments, it was possible to retrieve that data from Prolific Academic for Experiments 3, 4, and 5. We updated the Participants sections of these three experiments with the specific percentages of self-reported ethnicity. In Experiments 3-5, the majority of participants identified as being of White ethnicity.

Experiment 3

(…) Data about participants’ self-reported ethnicity was retrieved from Prolific Academic and revealed that most participants identified themselves as White (White = 70.4%, Black = 22.8%, Mixed = 4.4%, Asian = 1%, Other = 1.5%; according to the ethnicity categories provided by Prolific). (…)

Experiment 4

(…) Data about participants’ self-reported ethnicity was retrieved from Prolific Academic and revealed that most participants identified themselves as White (White = 78.5%, Black = 8.1%, Mixed = 8.1%, Asian = 3%, Other = 2.2%; according to the ethnicity categories provided by Prolific). (…)

Experiment 5

(…) Data about participants’ self-reported ethnicity was retrieved from Prolific Academic and revealed that most participants identified themselves as White (White = 63.6%, Black = 21.6%, Mixed = 9.8%, Asian = 2.1%, Other = 3%; according to the ethnicity categories provided by Prolific). (…)

**Comment 7.**

*Results*

* *Please indicate explicitly in the text that data were aggregated by observer (unless you change the analysis to a more appropriate mixed effects model). Although you didn’t explicitly do this in the code, apex does this for you (e.g., try explicitly calculating the mean for the 4 groups for each person — the ANOVA results will be identical)*
* *An interaction plot of the trustworthiness ratings for the 4 groups would be useful for interpretations here*

*I very strongly recommend using mixed effects model to analyse these data instead of an ANOVA. This will account for the fact that the face stimuli are sampled, in the same way that observers are sampled from a larger population that you wish to generalise to. This is additionally important because of the relatively small number of faces and imbalance of face characteristics between the groups. Otherwise, you can only make conclusions that generalise to the population of observers, but not ones that generalise beyond the sample of 24 stimulus faces you used.*

*However, this will require you to add Face ID labels to each trial. It is currently not possible to determine which face was shown on each trial, or I would have run this analysis for you to demo. See DeBruine & Barr (2021, AMPPS) for examples of a cross-classified mixed effects model similar to your study.*

**Response:**

We followed the suggestion of the Reviewer and re-analyzed all data accordingly. The main conclusions remain.

**Comment 8.**

*Code*

* *The dependency files setup.R and helpers.R are not included in the shared code.*
* *I could figure out most of the package dependencies, but couldn’t run the function interactionMeans() — I found compute\_alpha() in the Experiment 3 helpers.R file*
* *I could run the rest of the code, but did not check whether the values match the manuscript, as I will be recommending a different analysis, but it really helps me do reproducibility code check if you make it very clear how the analyses in the code correspond to the numbers in the results section. A great way to do this is to use sprintf() or glue() or inline R to create the text of the results section with inserted values from code.*
* *The data processing code is very hard to understand and needs comments (e.g., I can’t figure out why you created and add the column man\_check, as it’s identical to the existing Check column)*
* *FYI: the code from lines 55-79 could be condensed into
test=which( experiment\_l$stimulus %in% c( ‘A\_3\_1’, …rest of the values…, ‘D\_2\_2’) )*

**Response:**

We thank the reviewer for their detailed and rigorous review of our code. Following your suggestions, we aimed to clarify the code as best we can. Please note that the code of Experiments 1 and 2 was written by a different coder than the code in Experiments 3, 4, and 5. Translating code takes a considerable amount of additional time to be invested so we made a decision to minimize changes to this “legacy” code (of experiments 1 and 2), in view of the time we had to submit the revision. We will continue working on tidying the R code at OSF after re-submission, but will avoid any changes that break the code of Experiments 1 and 2. The differences in coding style did not affect the main results, and reflect only different stylistic approaches to data processing.

**Comment 9.**

*Experiment 2*

*Methods*

* *Same methodological critiques as exp 1*

*Results/Code*

* *Same advice as Exp 1 about mixed models*
* *Same problem with missing helper files*
* *I could run everything apart from the code that used interactionMeans()*
* *I am not really sure what that cronbach’s alpha is doing. You seem to be averaging the ratings over the six faces in each unique value of participant\_nr, story, attractiveness, trustworthiness, likert\_scale, and label, and then calculating the Alph for the 8 means per participant grouped by likert\_scale (judgement) and story (i.e., each combo of att/trust/label). This could use more clarity with comments in the code, as it’s not discussed in the manuscript at all. (Also relevant to Exp 1)*

**Response:**

We would like to refer the Reviewer to our previous responses. We made necessary changes where needed.

**Comment 10.**

*Experiment 3*

*P. 14 “Experiment 3 aims to replicate the findings of Experiment 2 with sufficient statistical power to detect a potentially smaller (i.e., d = 0.4) moderation of the bias towards computer-generated faces by background story.“*

*With 60 subjects, you have 80% power to detect d = 0.36 in a paired design, so Experiments 1 and 2 were already sufficiently powered for this.*

**Response:**

Experiment 3 was powered for investigating the interaction between the spin of the cover story and the label effect. We feel, however, that the Introduction of Experiment 3 does not do a good job in outlining its different aims. Hence, we rewrote it (p. 12):

“(…) Second, power calculations were now geared to specifically test an interaction between the effect of label and the background story. (…)”

**Comment 11.**

*Methods*

*I commend you for including a preregistration (*[*https://osf.io/w4bca*](https://osf.io/w4bca)*). It is very clear and detailed (seriously, I’m impressed!). This isn’t always the case, in my experience, so please include a clearer summary of what was preregistered and what deviations there were (e.g., the adjustment to the minimum time allowed to be spent on the story screen mentioned in the analysis script). Quite often, I see papers claim to be preregistered in a way that insinuates everything was specified, and after looking at the registration document find out they only roughly specified half the hypotheses and said they’d be analysed “with anova” without further details. Make sure a reader knows how detailed your pre-reg is in methods and analyses (and then they can get more details at the link if needed).*

**Response:**

We thank you for your compliment and suggestion. We took your suggestion into consideration while revising the manuscript.

**Comment 12.**

*P. 16 “The previously used 24 pictures (6 per face category; see Experiment 1) were complemented with 18 faces from the same database per category, adding up to a total of 96 faces.“*

*Please include the face image IDs for these extra faces, as well as their values for the category ratings.*

**Response:**

We added an additional table to the supplemental materials in OSF that reports stimulus IDs and respective Attractiveness and Trustworthiness CFD ratings for the stimulus pool used in Experiment 3.

**Comment 13.**

*Results*

*I’d also recommend adding a cross-classified mixed effects model for this analysis. Instead of calculating d’ for the memory data, you can use a binomial glmer. This will account for the randomising of face stimulus identities between subjects, as well as for stimulus sampling.   As the analyses here are pre-registered, you should definitely report the original planned analysis and compare them with the mixed effects analyses as robustness checks.*

**Response:**

Thank you for your insightful suggestion. We additionally ran the generalized mixed model with accuracy as the DV. The results entirely converged with the d-prime analysis using ANOVA or LMMs, showing no effect of Story nor a difference in accuracy between CG stimuli and Natural stimuli. To clarify why we just referred to a mixed model, these were run to replace the previous ANOVA analyses to keep consistency with the replacement of ANOVAs with LMM made throughout the manuscript revision. The LMMs had d-prime or criterion as DVs and additionally take into account stimulus variability (model <- lmer(dprime OR criterion ~ story\*label + (1 + label | id) + (1 | cfd\_image)…). This model specification was the same employed with the GLMM, using accuracy as the DV. We now report this alternative analysis and results in a new footnote (p. 19, Footnote 2):

Footnote 2

“Following a suggestion by a reviewer, we conducted an alternative, non-preregistered, generalized mixed model analysis to account for the randomizing of face stimulus identities between subjects, and for the stimulus sampling. The model was specified similarly to the d’ linear mixed models, with the exception that accuracy in the memory task was now the dependent variable (coded as 1= correctly identified stimulus as old or new, 0 = incorrectly identified stimulus as old or new). The results (reported on the logit scale) entirely converged with those obtained for the d’ analysis, showing no significant difference in accuracy between stimuli that had been labeled as natural and computer-generated (*bComputer-Generated\_Label* = -0.08, *SE* = 0.16, *p* = .63, Label reference level: natural; odds ratio (natural / computer-generated) = 0.95, *SE* = 0.11, *p* = .91).”

Finally, we now also clarify the deviations to the pre-registered analyses in the Data Analysis subsection (p.17):

“These analyses, following a suggestion by a reviewer, deviated from the originally pre-registered ANOVA analyses and improve upon them by taking into account more sources of variability in the data. The results of the pre-registered ANOVAs converged entirely with the results of the linear mixed models.

An identical clarification was added to the Data analysis subsection of Experiment 4 (p. 27). Experiment 5 pre-registered LMM analyses, so it did not require any clarification of ANOVA to LMM changes.

**Comment 14.**

*Code*

*Experiment 3 needs a README to tell the user what order to run the scripts in. I assumed Anonymization, Cleaning, then Analysis. Obviously, Data Anonymization won’t have the raw data shared, so please also explain this explicitly in the README.*

*The Data Cleaning script has very good comments/text and is much easier to follow than the scripts for Exp 1 and 2.*

* *Line 44 reads the data from the directory “data”, but this is actually called “anonymized raw data” in your shared materials.*
* *This script ran fine after fixing this*
* *I skimmed the code for sense and didn’t see anything obviously wrong, but didn’t do line-by-line code check*

**Response:**

We thank you for your compliment and suggestion. We will add to the OSF repository a README file explaining the nature of the data and provide instructions on the order of analyses, where needed.

**Comment 15.**

*The Data Analysis code ran fine without changes 😃*

* *The long narrative comments (e.g., lines 38-56) could be moved out of the code chunks into the markdown section of the scripts. This usually makes the HTML output easier to read. This is just stylistic advice, but I tend to put scientific narrative in the markdown and use code comments only for specifically code-relevant notes.*
* *The alpha code produces a LOT of messages: “In smc, smcs < 0 were set to .0”; it might be useful to hide those and add a comment explaining why.*
* *In the Data Processing section, it’s not 100% clear why you drop\_na() and loose about 6000 observations, but these are all observations with NA in the label column. It would be useful if this section had a little more narrative about the processing.*
* *I didn’t have capacity to do line-by-line code check*

**Response:**

We apologize for this and clarified the code as best we can.

**Comment 16.**

*General*

*Figure 1 implies the Likert ratings were a sliding scale, but the data show values as integers. Also, the figure doesn’t match the methods text on Page 9: “Below each face, two 7-point Likert scales were presented. One for attractiveness and one for trustworthiness. For both scales the right side was labeled with “absolutely not” (1) and the right side with “extremely” (7). “*

**Response:**

We apologize for this and removed this figure.

**Comment 17.**

*I’m a little confused by Figure C. The y-axis is labelled observations, which I would normally assume are integers, but the histograms start at 0.1 and go as high as 0.6, so I assume this is either some type of density plot or an artifact of plotting the horizontal box plots on the same plot. The x-axis is perceived trustworthiness, with labels 2, 4 and 6, but the histogram bars seem to be 0.5 in width, so are these plotting mean rating scores or individual observation ratings?*

**Response:**

We apologize for this. We decided to drop this figure and create new ones to accommodate the additional studies and took your comment into consideration while generating new plots. The same information about the label effects is now shown in a new Figure 3 showing raincloud plots for all five experiments. We hope the axes of the plot will now be clearer to the reader.

**Comment 18.**

*I spent 20 minutes installing new packages (some dependencies had to be installed from source), so make sure you’re only calling what you need from the setup.R file. Thanks for including the install code for non-CRAN packages 😃*

*FYI: You can resolve functions conflicts by loading the package with default functions last (e.g., load tidyverse last to use dplyr functions as defaults). This avoids functions in the global environment. (But is more a stylistic choice and not a problem.)*

**Response:**

We apologize for this and optimized the installation routine. Thank you for the useful tips, it is always a pleasure to learn new things while revising a manuscript.

**Reviewer 2**

**General comment:**

*The manuscript “Faces Merely Labelled as Artificial are Trusted Less” investigates whether the relationship between AI and people is in the mind of the beholder. In three experiments, the authors showed that participants perceived natural faces that were merely labelled as artificial as less trustworthy than natural faces labelled as natural. This is an interesting and relevant topic considering the increasingly pervasive presence of computer-generated faces in our everyday life. These results open a new question: why do people trust computer-generated faces less than natural faces? Previous research had hypothesised that people are less expert in inferring traits from artificial faces because we do not see them quite as often as natural faces, but this study showed that it is possible to modulate people’s judgment also when facial features are natural. A (pre-registered) working hypothesis made by the authors was that artificial faces would have been considered as part of an outgroup. This was measured using a memory task, but the null hypothesis could not be rejected.*

*I generally liked the manuscript which reads very nicely and it is quite clear in every section. I very much appreciate the pre-registration for study 3 and the fact that all scripts and data are shared already. The manuscript is well-written, relevant and opens new interesting questions about the relationship between AI and humans.*

*I have a few major clarifications and some minor suggestions to hopefully improve the readability of the manuscript.*

**Response:**

We thank the Reviewer for this positive assessment.

**Comment 1.** *Major 1:
As I said, I very much enjoyed reading the manuscript, but I feel the results would have been easier to grasp using more appropriate figures. As it is, there is only one figure and the subpanels are not even properly referred to, unless I missed it. It would be very useful to see the results of the ANOVAs, and possibly to show individual participants to have an idea of the sample distributions. The description of the results of the ANOVAs is great, but some readers (like me) would prefer to see them also in a plot.*

**Response:**

We apologize for this and, more optimal figures are now presented. Please note that we no longer used RM ANOVAs. We paste below all the new Figures.

**Figure 1**

*Illustration of the Rating Task and Stimuli of Experiments 1, 2, 3, and 5*



*Note.* Panel A: Example of the face rating task in Experiments 1, 2, 3, and 5. Only one face was shown at each trial, paired with only one of the labels. Panel B: Example of stimuli extracted from Experiment 3 that were rated, on average, as high (top row) or low (bottom row) in trustworthiness.

**Figure 2**

*Distributions of Believability Ratings for All Experiments*



**Figure 3**

*Raincloud Plots of the Main Effects of Label in All Experiments*



**Figure 4**

*Experiment 4 Task Overview*



**Figure 5**

*Average Effect Size of the Label Effect Across All Experiments*



**Comment 2.** *Major 2:
I was wondering whether the type of story that you used could have somehow revealed the aim of the study. Did people understand that you wanted to find a difference between artificially generated faces and natural faces for the two stimulus characteristics? I am not saying that this is the case, but I wanted to get your opinion on this. It is true that you don’t see the same effect on attractiveness for experiment 1 and 2, and this could be a good sign that your experimental manipulation was specific to trustworthiness. However, attractiveness could be more difficult to manipulate without manipulating the perceptual aspect of the stimuli. At the same time, you do see an effect on attractiveness in experiment 3. I would just like to hear if you think the possibility of this type of bias can be ruled out, and eventually add a sentence or two in the discussions.*

**Response:**

We thank the Reviewer for this thoughtful comment. The use of explicit ratings is always prone to demand and framing effects. Similar suggestions were made by the Editor and Reviewer 1. Accordingly, we added two experiments with the aim of further testing the boundary conditions of the label effect. Of interest here, are the results of Experiment 5 (p. 28), which suggest that the label effect is absent when participants only have one type of face to rate. We believe this is an important finding, which has implications for research on Humans and AI in general.

We consider these issues in the GD (pp. 34-5):

“The results of Experiments 1-3 are difficult to reconcile with our initial hypothesis that the faces labelled as being artificial are represented as an outgroup, which results in these faces to be rated as less trustworthy (Balas & Pacella, 2017). At the same time, Experiment 4 further confirmed that the label effect is originated at the early stages of impression formation without the ‘behavior’ of the face coming at play. Finally, Experiment 5 suggests that this impression formation depends on the presence of different category of faces, which may indicate that some comparison or contrast between these categories is needed for the label effect to occur. One possible account that may fit these different findings is that the label effect is not mediated by an outgroup bias but reflects a more general evaluative conditioning effect. Evaluative conditioning leads to a change in valence of a stimulus due to the pairing of that stimulus with another stimulus that is intrinsically negative or positive (see Hofmann et al., 2010 for a review). In the context of the current study, it is possible that labels referring to AI (e.g., computer-generated, artificial, synthetic) are sensed to be less positive compared to labels referring to real entities (e.g., human, natural, real). As such these labels may function as unconditioned stimuli, which bias attitudes towards the faces they are paired with. As a result, faces paired with labels referring to AI are perceived as being less positive and, more specially, less trustworthy. Although such account is speculative at this stage it indicates that future research is needed to further specify the processes underlying the difference in processing artificial and real faces and test the boundary conditions of these differences. For instance, recent studies reported that state-of-the-art synthetic faces that are undistinguishable from real faces elicit higher trustworthiness ratings (Nightingale & Farid, 2022).This bias seems to depend on the degree to which these faces are believed to be real (Tucciarelli et al., 2022). Such finding corroborates with the current results by indicating the importance of beliefs and attitudes, which were explicitly manipulated in the current study by using labels.

Although we tested several boundary conditions of the label effect, the present study was restricted to the use of explicit ratings. Such ratings are assumed to reflect higher-order processes of deliberate reasoning, which may be affected by task demands. In contrast, attitudes may also result from automatic processes that occur spontaneously and outside of people’s awareness or control (Bargh & Williams, 2006; Moors & De Houwer, 2006). Such ‘implicit’ attitudes are typically inferred from people’s performance on response latency measures, such as the Implicit Association Test (IAT; Greenwald et al., 1998) or sequential priming tasks (Fazio et al., 1995; Wittenbrink et al., 1997). Importantly, explicit and implicit attitudes often do converge. For instance, Van Dessel et al. (2020) demonstrated that explicit attitudes towards social groups could be changes, but not implicit attitudes. Previous research already started to explore implicit attitudes in the context of social robotics (e.g., Diana et al., 2022; Erel et al., 2019) and question remains whether the label effect can be generalized when using more implicit measures.”

**Comment 3.** *Major 3:
I was initially suspicious about your choice of using the ANOVA to analyse your data, because you are essentially only treating the participants as a random effect and the stimuli as a fixed effect which could inflate Type I error (see Judd et al., 2012, J Personality and Social Psychology). However, due to the nature of your stimuli (i.e., they are always the same faces) and the fact that you counterbalanced the labels (computer-generated and natural) between participants, this might not be a problem, as any effect due to the stimuli should cancel out. In any case, using the ANOVA makes the analysis a bit complicated (e.g., in experiment 2 and 3 you have a 4 factors design) and you cannot control directly for covariates (e.g., gender, age, attractiveness judgment). Wouldn’t have been just simpler and more appropriate to use mixed-effects models here? Can you otherwise rule out that the effects that you see are not due to differences in gender, age, or attractiveness judgments?*

**Response:**

In line with the comments made by you and the Editor and Reviewer 1, we decided to reanalyze our data by using linear mixed effect models. The main conclusions of our study remain the same.

**Comment 4.** *Major 4:
I was also suspicious about the choice of using ANOVA because your dependent variable is not continuous but it is ordinal. It is probably common to use ANOVA in this situations, but this does not justify using it. Have you checked that the assumptions of the ANOVA are met?*

**Response:**

As indicated by our previous response we now solely use linear mixed effect models.

**Comment 5.** *Minor 1:
Panel D seems a nice figure, but I don’t think it is referred to in the manuscript.*

**Response:**

We thank you for the compliment and for alerting us to the lack of reference to it. However, in light of the current revisions, we have decided to remove the original figure and create new figures that are better tailored for the latest version of the manuscript. In this respect, we have now created a small final section entitled “Meta-analysis” placed just before the GD, where we explain how we calculated the average effect size of the Label effect, and included a new Figure 5, showing an updated version of the forest plot including the effect sizes for all experiments. We believe this new section and figure add value to the manuscript by directly offering an average effect size to the reader.

**Comment 6.** *Minor 2:
“Attractiveness judgements are mainly based on a global affective response that requires minimal inferential activity (e.g., Zajonc, 1980) and offers a benchmark for more sophisticated judgments such as of trustworthiness (see also Willis & Todorov, 2006).”
What do you mean with “offers a benchmark” here? And why did you choose attractiveness in this experiment?*

**Response:**

We thank you for your comment and apologize for not being clear. As you already clarify in one of the major points (Major 2), attractiveness was included to assess the extent to which the findings would be exclusive to trustworthiness (and not a more general valence dimension/ general attitude). We added an additional sentence to the Introduction to better explain our decision (p.5):

“In other words, the inclusion of attractiveness judgments helps assess the extent to which any effect of labelling is exclusive to a trustworthiness dimension, as opposed to any other dimension strongly related to general valence.”

**Comment 7.** *Minor 3:
Related to Minor 2. Why you used the attractiveness associated to each face instead the attractiveness estimated by the participants?*

**Response:**

We opted to use the original ratings of the face database as these were obtained with larger samples and seemed like the most efficient approach to build the initial experiments (1 and 2). It is also a common practice in the face research literature to use pre-validated stimuli to manipulate trait perceptions (via the trait signal communicated by the configuration of the stimulus’ features) in experiments where these perceptions are not focal but a means to any other research goal. To further clarify, the participants did rate the attractiveness of the stimuli in Exps. 1-3, and it were these observed ratings that were analyzed. The reported main effect of face attractiveness category for attractiveness ratings clarifies if the manipulation of attractiveness was successful, by showing that faces that were classified as high or low in attractiveness based on the original database ratings, were indeed perceived as such in our experiments. This was indeed what we observed in our results (Exps. 1-3).

**Comment 8.** *Minor 4:
“In each category, faces were randomly divided in two sets of 3 faces. These sets were either labelled as being natural or computer-generated faces. This labelling was counterbalanced over participants.”
Could you please clarify whether there were always the same 3 faces in each set?*

**Response:**

In Experiments 1 and 2, the face identities allocated to a Face attractiveness/trustworthiness category were the same. However, in Experiment 3 the three faces identities of each category were randomly drawn each time from a larger stimulus pool constrained to match the particular category. For example, in Experiment 3 the three faces high in attractiveness and low in trustworthiness were sampled from a set of 10 faces which were all high in attractiveness and low in trustworthiness.

**Comment 9.** *Minor 5:
“Participants were either first presented with the set of faces labelled as natural or with the set of faces labeled as computer-generated. Each set consisted of 12 faces (3 faces per category). Within each set faces were presented in a random order, one at a time in the middle of the screen.”
Could you please clarify whether also the category sets were randomly presented?*

**Response:**

We apologize for not having been clear. Indeed, that was the case. We added a sentence in the manuscript to clarify (p.7):

“Each set consisted of 12 faces (3 faces per category). Face categories were randomized within a block such that any face category could change to another at every trial. Within each set, faces were presented in a random order one at a time in the middle of the screen.”

**Comment 10.** *Minor 6:
“Below each face, two 7-point Likert scales were presented. One for attractiveness
and one for trustworthiness. For both scales the right side was labeled with “absolutely not” (1) and the right side with “extremely” (7).”
From Figure 1A, it seems that the scale was continuous, not a Liker scale, and the labels were “Not at all” to “Extremely”. Which ones are correct?*

**Response:**

We apologize for this and removed this figure. The anchors in the figure implied the exact same anchors were used across all experiments. However, there were slightly different wordings for the Low scale extreme in the first experiments (Absolutely not instead of Not at all). W enow have a new Figure 1 where we clarify the specific scale anchors in each experiment.

**Comment 11.** *Minor 8:
“Participants were also informed about the nature of the upcoming set of faces (i.e., natural or computer-generated).”
When was this presented? Was it before each face? Before each set? Always present? Please clarify.*

**Response: (Please note that the revision list of Reviewer #2 seemed to skip Minor point 7)**

We thank you for noticing this was unclear and have now clarified it in the manuscript (p.7):

“At the beginning of each block, participants were also informed about the nature of the upcoming set of faces (i.e., natural, or computer-generated).”

**Comment 12.** *Minor 9:
Regarding Figure 1B, could you show the percentage rather than the absolute values? I think it would facilitate the comparisons between experiments. How do you explain the fact that many people (especially in experiment 3) did not believe that the faces were artificial, but you can still see an effect on trustworthiness? This might be related with Major 2, meaning that the aim of the study was evident to participants.*

**Response:**

We have conducted a new experiment using a between-subjects design for the label condition (Experiment 5). The results revealed no label effect this time, which may be related with the lower scores of believability of the label in the Computer-generated label condition, compared to the Natural condition. Finally, we removed that figure and created a new Figure 2 showing how participants (%) were distributed across the believability scale within each experiment (Exps. 1-5).

**Editor First Decision**: Revise & Resubmit

Feb 8, 2023

Dear Manuel Oliveira,

I have now read your revised manuscript. I appreciate your careful attention to the concerns the reviewers and I raised. I am happy to provisionally accept your manuscript for submission. However, I found a few small things I would like you to address.

Minor revisions:

* The General Discussion says: “In Experiments 1-4, we observed that natural faces merely labeled as being artificial were judged to be less trustworthy.” This implies that the effect was significant in those studies, but this was not the case for Experiment 2. Throughout the manuscript, the results of Experiment 2 should be treated as though they did not support the main hypothesis, as they did not provide evidence that would warrant rejecting the null.
* The manuscript should list exact p-values, except when p < .001.
* The title – “Faces merely labeled as artificial are trusted less” – feels a bit misleading in light of the results of Experiment 5. I think it would be more accurate to change the title to “Faces merely labeled as artificial are trusted less when contrasted with faces labeled as natural” or something similar.

I look forward to receiving your final revision and accepting it for publication in Collabra: Psychology.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Alexa Tullett

**Author Response**
Feb 14, 2023

Dear Professor Alexa Tullet,

We want to start by expressing our enthusiasm with your decision to provisionally accept our manuscript. With great pleasure, we submit for your consideration our revised article, now entitled “Are natural faces merely labelled as artificial trusted less?” following your feedback. Below, we briefly summarize our responses to the minor points raised in your latest review.

Again, we want to thank you for providing us with valuable feedback that allowed us to strengthen our paper. We thank you for the chance to resubmit our work to *Collabra: Psychology* and look forward to hearing back from you regarding the suitability of this final revision for publication in *Collabra: Psychology*.

With best regards,

And on behalf of all the co-authors,

Dr. Baptist Liefooghe

Dr. Manuel Oliveira

Dr. Ruud Hortensius

**Editor Final Decision:** Accept

Mar 1, 2023

Dear Manuel Oliveira,

I have now had a chance to read over your manuscript “Are natural faces merely labelled as artificial trusted less?”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers and I raised. I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point.

You will be receiving separate correspondence regarding any production and technical comments, data deposits, as well as publication charges. We work with the Copyright Clearance Center to process any applicable APC charges. Please note that your APC transaction must be completed before your article gets published.

You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely,
Alexa Tullett