**Revision**

**Editor and Reviewer comments:**

I have received three thoughtful reviews of the manuscript that you submitted to Acta Psychologica (ACTPSY). The reviews are appended below. The reviewers acknowledge that the paper addresses interesting research questions, and they applaud the aspirations represented in the paper. However, the reviewers have also pointed out several weaknesses of your manuscript which need to be addressed before it can be considered for publication by ACTPSY. If you decide to revise the work, please carefully consider all reviewer comments.

As soon as the authors deal with the points evidenced here (or provide a sound explanation of why they disagree with them), I'd be happy to see the paper published !

**Reviewer #1**

The study addresses a topic of extreme relevance. I can clearly see the theoretical and applied justification for research of this nature. In general terms, the article is well written and the methodology seems on the right track. However, I consider that more information/development is needed on certain relevant aspects of the research.

1. Intro

While the rationale for choosing attractiveness as a variable is well elaborated, there is an important contrast when it comes to elaborating the choice of the variable’s trustworthiness and familiarity. I understand that attractiveness is the main variable, however, the link between attractiveness and trustworthiness/familiarity is mentioned very briefly and seems to be underdeveloped. Readers that are nor familiar with literature on face perception may find the inclusion of those variables somewhat random. It requires better justification.

To-do:

1. Develop **link** between **trustworthy/familiarity** and **attractiveness**

2. **Justify** the **rationale** behind this choice

**Attractiveness:**

*Similarly, trustworthiness ratings appear to be heavily influenced by prior beliefs about the nature of the stimuli and their perceived artificiality (Wang & Nishida, 2024). Rated artificiality significantly affects trustworthiness ratings of synthetic faces. When participants are informed that images were synthesized by AI, artificiality is negatively correlated with trustworthiness ratings (Wang & Nishida, 2024). Additionally, real faces labeled as AI-generated receive lower trust ratings (Liefooghe et al., 2022). Conversely, when participants are unaware of the AI nature of the stimuli, trust ratings for AI-generated faces increase (Nightingale & Farid, 2021). This suggests that perceptions of authenticity and prior knowledge about the nature of faces play critical roles in determining trustworthiness ratings. H*owever, whether trustworthiness might influence reality beliefs is also underdeveloped.*.*

We expect a similar relationship with trustworthiness ratings …. *as well as findings on trustworthiness judgments in AI-generated faces (Nightingale & Farid, 2021).*

**Familiarity:**

Average faces are perceived as more attractive and familiar?

2. Method

In line with what was mentioned above, I find that certain methodological decisions should be better justified.

Notably, more information about face stimuli seems mandatory to me. Beyond the database and emotional expression of the faces, it is difficult to evaluate the suitability of the stimuli in the current version of the manuscript. For instance, what were the selection criteria? How many faces were male/female? Why was the number of faces selected (109)? Were dimensional properties ​​inherent to face stimuli (e.g., arousal, dominance) controlled? More information is needed.

To-do:

1. what is the **selection criteria** for the stimuli?

2. **Number** of male/female faces

3. **Why** 109 faces?

4. **dimensional** properties **controlled**?

There were **no selection** criteria for the stimuli, other than using a validated database with a large and diverse set of neutral faces. We decided to opt for the dataset including only neutral faces to avoid any **avoid any potential confounding variables** such as that of affect. The American Multiracial Face Database comprises of a total of **110** unique faces, each (except one) posing with a neutral or smiling expression. Subsequently, 109 photographs exist in the database of individuals with a neutral expression, while 110 exist of individuals with a smiling expression. The selected database consists of **89 women and 20 men**, with neutral faces.

I don't know if there is a word count limit. Otherwise, I find a certain lack of theoretical justification when introducing the various scales used. More information on other methodological decisions is also relevant. For example, what is the reason for the decision to present the faces for 500 ms? Is there any reference in the literature or was it an arbitrary decision by the authors? The decision to use two scales to measure attractiveness seems correct to me, however, it could also be explained better. For instance, does it partially respond to the need to have a subjective/objective measure of the construct?

To – do:

1. More theoretical **justification** for the **scales**

2. **Why** are faces presented for **500ms**?

Based on pilot studies, this duration was found to offer a sufficient perceptual window for individuals to make decisions. Previous findings also indicated that judgments reached a stable average level with the 500-ms exposure, with no significant changes observed upon increasing it, except for higher confidence levels (Willis & Todorov, 2006).

3. **Explain** the use of 2 **attractiveness** **scales** (is it a subjective/ objective measure of the attractiveness construct?)

**Reviewer #2**

This paper aimed to unveil some of the underlying factors that drive reality judgements of face images. The authors presented participants with real human faces which were initially judged according to different traits. Participants were presented with the images a second time and led to believe that some of the faces were AI-generated—when they were actually all human—and were asked to rate the realness of each face. The authors also assessed the role of individual differences in traits.

This paper engages with a very interesting and increasingly important topic. The research design was well-conceived, and it was clear that a lot of thought went into the choice of measures (e.g., separating out general attractiveness and physical beauty).

Overall, the current study is of a high standard; however, I do have some suggestions for improving the paper:

1. The literature review presented a good overview of relevant research and theoretical models. Research on AI is rapidly moving, so it is difficult to keep up to date with the literature. I stumbled upon a paper which had very similar objectives to the one presented here but I believe it was published after the current paper was submitted, I am not necessarily recommending to include this paper, but it is worth reading and perhaps integrating into the introduction section if it seems relevant.

Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A., Krumhuber, E. G., & Dawel, A. (2023). AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. Psychological Science, 09567976231207095.

To-do:

**1. Read** this paper

2. Include in intro if relevant

2. I would have liked to have seen far more justification and rationale for the key factors that were investigated in this study. I understand that many of the hypotheses and research questions are exploratory and there may not be a great deal of past research to draw upon; however, it is not clear to me why these factors were chosen. Although the predictions around attractiveness were well-justified, I would still like to see a bit more elaboration on the affective reality theory as it was a key justification for the main focus of the paper yet was only mentioned once. Similarly, I was a bit surprised to see traits such as narcissism and honesty-humility were included as a key factor in the study. Although some studies were cited to support this, it is important to outline the proposed mechanisms underlying this.

To-do:

1. Give more **justification** and **rationale** for the choice of **key factors**

2. **Elaborate** the **effective reality theory** throughout the intro

3. Outline proposed **mechanisms** **underlying** the choice **of honest-humility and narcissism** traits as key factors?

3. The overarching design of the study was effectively designed to address the research questions. However, I found much of the finer details of the method section a bit unclear. Specifically, there were a lot of details that were briefly mentioned but not clarified. I have listed some points below that warrant clarification:

a. The AMFD was an appropriate database to use. However, could you please outline the gender and race composition of the stimuli used?

To-do:

1. Gender and race **composition** of the stimuli

The stimuli comprised of 89 women and 20 men. All the participants reported having either a racial background consisting of two racial groups (81) or racial background consisting of three or more racial groups (29), where 33% self-reported as Asian/White, 22% as Latinx/White, 11% as Asian/Latinx, 6% as White/Middle Eastern, 5% as Black/White, and 5% as Asian/Middle Eastern.

b. It is not clear to me why stimuli were only shown for 500ms. We can inform impressions of faces very quickly, but in the real world we wouldn't be faced with AI detection under time constraints. Are you able to justify this decision (e.g., to reduce the overall duration of the study)?

To-do:

1. **Why** was stimuli presented only for **500ms**?

Based on pilot studies, this duration was found to offer a sufficient perceptual window for individuals to make decisions. Previous findings also indicated that judgments reached a stable average level with the 500-ms exposure, with no significant changes observed upon increasing it, except for higher confidence levels (Willis & Todorov, 2006).

c. I encourage you to include a 'design' section within the method.

To-do:

1. Add **design section**

This study makes use of a within-subjects design with repeated measures. The main independent variables (IV) investigated are subjects’ self-perceived attractiveness and the perceived realness of the stimuli, which is manipulated prior to the start of the study, by informing participants that half of the faces presented are artificially generated.

d. Can you please elaborate on the following: "we also planned to do between-participants analyses". what between-participants analyses were performed?

To-do:

1. what **between-participants** analyses were performed?

e. I would like to have seen more details around data screening. For instance, what % of incomplete responses wanted removal of participants? Were any data imputation methods used for missing responses? If not, could this be done.

To-do:

1. develop **data screening** procedure

Missing data was prevented by using **a forced-choice questionnaire,** in which participants were not able to proceed to the following display until a response was made for all the items in the survey.

f. I agree that it is important to focus on sexual-orientation relevant stimuli given the focus on attractiveness judgements. However, what did you do if participants did not provide their sexual orientation? Moreover, which stimuli did participants view if they were attracted to more than one gender, such as bisexual; did they view all stimuli?

To-do:

1. what was **done when** participants did **not provide sexual orientation**?

2. If they were **bisexual**? Did they **view all the stimuli**?

Participants who did not provide sexual orientation were **excluded** from the analysis, and those attracted to more than one gender were **presented with all the** stimuli.

g. Could you please run reliability analyses on the scales used in the study and report the Cronbachs Alpha

To-do:

1. Report **reliability analysis** on the scales used

Scales:

1. Mini-IPIP6:

a) extraversion’ raw\_alpha 0.82  
 b) Conscientiousness’ raw\_alpha 0.73  
 c) Neuroticism’ raw\_alpha 0.59  
 d) Openness’ raw\_alpha 0.73  
 e) Honest-Humility’ raw\_alpha 0.7  
 f) Agreeableness’ raw\_alpha 0.82

2. SIAS-6 and SPS-6

a) Social Interaction’ raw\_alpha 0.84 (SIAS)  
 b) Social phobia’ raw\_alpha 0.87 (SPS)

3. Expectations about AI’s raw\_alpha 0.61  
  
4. GAAIS’s   
 a) positive’ raw\_alpha 0.84  
 b) negative’ raw\_alpha 0.59

5. FFNI-BF (high-order dimensions)  
 a) Agentic narcissism’ raw\_alpha 0.78  
 b) Antagonistic narcissism’ raw\_alpha 0.83  
 c) Neurotic Narcissism’ raw\_alpha 0.79  
  
6. R-GPTS  
 a) Reference’ raw\_alpha 0.85  
 b) Prosecution’ raw\_alpha 0.93  
  
7. IUS-12  
 a) Prospective anxiety’ raw alpha 0.8  
 b) Inhibitory anxiety 0.82  
  
7. Self-rated attractiveness’ raw\_alpha 0.94

4. It is great that the authors have deeply engaged with open science practices; this is a clear strength of the paper. I understand that the preregistration and other materials were blinded due to the need to maintain anonymity during the review process; however, it would be immensely helpful to review this paper if I were able to access these links. It is possible to share open science links without it being identifiable: <https://help.osf.io/article/155-create-a-view-only-link-for-a-registration>

To-do:

1. Create **view-only link** for registration paper

5. Similar to the previous point, it appears that the paper has been preregistered (which is great!); however, it important to outline and justify any deviations to the preregistration (perhaps as an appendix). It is completely fine if there are deviations to the preregistration, but it is important to be transparent about these deviations and explain them.

To-do:

1. **Outline** and **justify** any **deviations** from the **preregistration**

In the first part of the study participants also reported self-rated attractiveness which was mentioned in the measured variables section of the preregistration but not clarified as being assessed in this part.

Wording of the attractiveness, trustworthiness and familiarity changed slightly, by removing the “Assuming the face you saw was of a real individual…”. All questions were changed to “I find this person…” or “This person ….”

The second scale for attractiveness (beauty) was added to the study but not mentioned in the preregistration as something being assessed in the measured variables section.

In the pre-registration the ratio presented as real or generated, in the last part of the study, was 50-50. However, in the actual study this changed to “about half”.

**Data-analysis:**

Belief ratings were analysed using logistic mixed models which modelled the probability of assigning a face to real or fake. This was not included in the analysis plan of the preregistration.

Confidence ratings as well as attractiveness, trustworthiness, beauty, familiarity were modelled using mixed beta regressions

6. Mixed effects models were appropriate given the nature of the data. However, I found the results section very hard to follow. Namely, the model specifications were not always clear to me. It might be helpful to list out how the models were constructed (e.g., what were entered as fixed and random effects). Similarly, were any of the variables nested (e.g., based on sexual orientation/ gender)?

To-do:

1. Clarify result section

2. List out **models constructed** (what were the fixed and random effects?)

3. **Nested variables**?

**1) Delay of stimulus re-exposure and presentation order**  
  
For models examining the influence of delay and presentation order (fixed effects) on participants' beliefs of reality, random intercepts for participant and stimulus were incorporated.

For models examining the influence of belief and delay, and belief and presentation order, as well as their interactions (fixed effects) on confidence judgments, delay and presentation order were nested within belief. Random intercepts and slopes for participants for belief and the nested variables were incorporated, along with random intercepts for stimuli.

**2) Determinants of Reality**

a) attractiveness

This model examines the influence of attractiveness, nested within sex, on beliefs, incorporating participant and stimulus as random intercepts. The fixed effects include sex, both linear and quadratic terms for attractiveness, and their interactions.

This model examines the influence of belief, sex, and attractiveness (including both linear and quadratic terms), as well as their interactions, on confidence judgments. The model specifies that sex is nested within belief, and attractiveness is nested within sex and belief. It incorporates random effects (intercepts) for participants and stimuli.

b) Beauty adjusted for trustworthiness and familiarity

This model examines how beliefs are influenced by sex and beauty (represented by both linear and quadratic terms), while adjusting for the effects of trustworthiness and familiarity. It includes random effects (intercepts) for participants and stimuli.

This model examines the influence of belief, sex, and beauty (including both linear and quadratic terms), while adjusting for the effects of trustworthiness and familiarity. It specifies that sex is nested within belief, and beauty is nested within sex and belief. The model includes random intercepts for participants and stimuli

c) Trustworthiness adjusted for beauty and familiarity

This model examines how beliefs are influenced by sex and trustworthiness (represented by both linear and quadratic terms), while adjusting for the effects of beauty and familiarity. It includes random effects (intercepts) for participants and stimuli.

This model examines the influence of belief, sex, and trustworthiness (including both linear and quadratic terms), while adjusting for the effects of beauty and familiarity. It specifies that sex is nested within belief, and beauty is nested within sex and belief. The model includes random intercepts for participants and stimuli

d) Familiarity adjusted for beauty and trustworthiness

This model examines how beliefs are influenced by sex and familiarity (represented by both linear and quadratic terms), while adjusting for the effects of beauty and trustworthiness. It includes random effects (intercepts) for participants and stimuli.

This model examines the influence of belief, sex, and familiarity (including both linear and quadratic terms), while adjusting for the effects of beauty and trustworthiness. It specifies that sex is nested within belief, and beauty is nested within sex and belief. The model includes random intercepts for participants and stimuli.

**3) Inter-Individual Correlates of Simulation Monitoring**

7. Related to the previous point, it was not clear to me why each key variable was run in a separate model/ analysis. That is, I would have assumed that it would be more effective to account for multiple variables within a single model, such as adding attractiveness, trust, familiarity etc as fixed effects within the one model rather than testing them all separately. My understanding is that this would account for the relative influence of the different effects and give a better idea of the interplay between them. You could also run likelihood ratio tests to compare different models to help us understand which factors are more impactful than others. I have only used mixed effects models in R a few times, so please don't hesitate to correct me if I am missing something obvious.

To-do:

1. why were **key variables** run in a **separate model/analysis**?

2. Run **likelihood** **ratio** **tests** to compare different models

8. Could possible confounding influences (e.g., exposure delay, presentation order), such as the issue of familiarity be partly mitigated by being entered into the models as random effects?

To-do:

1. Are **confounding** influences mitigated by being added as **random effects**?

9. The discussion section presented some good explanations for the findings. The authors also highlighted key limitations and did a good job explaining how these concerns were mitigated. However, my primary concern for the discussion, and the paper as a whole, was the sheer number of concepts that were investigated. It made the paper difficult to follow and many points were not explained in sufficient depth I suspect due to concerns around the word limit.

To-do:

1. Provide **further** **explanation** for the **findings**

**Reviewer #3**

The manuscript "Too Beautiful to be Fake: Attractive Faces are Less Likely to be Judged as Artificially Generated" addresses an issue of the utmost (and ever-increasing) relevance, namely the relationship between some features of face stimuli (attractiveness, beauty, trustworthiness, familiarity) or of the onlookers (e.g. their personality traits and attitudes toward AI) and the judgment of being real or fake (i.e. AI-generated).

As the authors themselves acknowledge, the paper does not provide any 'groundbreaking' insight in the matter. But since our aim in science is (and must be) understanding phenomena rather than winning click-baiting context, I am convinced that this paper deserves to be published (rather than fuelling the file-drawer problem). It will become a cornerstone upon which further studies could profitfully build upon (especially because "all the details, scripts and complimentary analyses are open-access", which may facilitate cumulative knowledge-building).

That being said, before endorsing publication, I recommend to deal with the following conceptual issue:

As the authors themselves acknowledge at the beginning of p. 22, a limitation of the study is that, due to the task instructions, participants expected that "about half of the faces were AI-generated and the other half real photos". Hence, while I take the balanced responses of participants (44% judged fake / 56% judged real) as prima facie evidence that task instructions were effective, I am less convinced that it speaks in favor of "it is to note that the paradigm did not instruct participants to balance their answers according to a certain distribution (e.g., 50-50), merely providing them a description of the dataset." (p. 17).

To put it simply, I think that, due to its design, the key findings of this study regard WHICH images (based on the 4 ratings) are more likely to be judged as real & by what kind of subjects, not WHETHER images are thought as real. Indeed, this second point has been strongly prompted by the instructions and has been proven by experiments tackling that issue more straightforwardly (e.g. Nightingale & Farid 2022; Miller et al. 2023).

Therefore, I suggest a small tweak in the narrative so as to emphasize the real contribution of this study in the relevant discussion sections (e.g. on p. 22, the following sentence might be rephrased: "it is to note that the paradigm did not instruct participants to balance their answers according to a certain distribution (e.g., 50-50), merely providing them a description of the dataset"

To-do:

1. **Rephrase** this sentence

please check and discuss the following study by Miller et al. (2023): AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones (https://journals.sagepub.com/doi/full/10.1177/09567976231207095)

To-do:

1. **Read**, and **add**, this **paper**

I also suggest a few amendments to the text:

1) Very minor issue, but I'm not entirely convinced by the very 1st sentence of the paper: "For the first time in human history, technology has enabled the creation of near-perfect simulations indistinguishable from reality". Indeed, I am not 100% convinced the real issue about current deepfake technology is that no technology before allowed to simulate reality. In a way, photographic manipulations are as old as photography itself (e.g. check the Wikipedia page of Oscar Gustave Rejlander, whose manipulated photos were also included in Darwin's book The Expression of Emotions in Man and Animals). In my opinion, what makes the current surge of synthetic media interesting and worrisome is that new technologies have made manipulations cheap, quick, and affordable: it takes but a few clicks to create synthetic faces with MidJourney VS the weeks of skilled work it took to recreate them in CGI.

I do not want to enforce this really minor point, but give it a think!

To-do:

1. **Rephrase** 1st sentence to highlight that it is not the first time that near-perfect simulation are indistinguishable from reality, but that nowadays is **cheaper, quicker and more affordable** to do so.

*Advancements in technology have now made it possible to create near-perfect simulations that are indistinguishable from reality. These simulations are more affordable, faster to produce, and more accessible than at any other time in human history.*

2) In the second paragraph, I found the sentence between brackets to be a tad too long: "While not all simulations have achieved perfect realism (Corvi et al., 2022; e.g., Computer Generated Images - CGI in movies or via recent algorithms such as GANs or diffusion models often include distortions or lack certain key details that makes them visually distinct from real images, McDonnell & Breidt, 2010), it". Consider breaking the sentences.

To-do:

1. **Break** the **sentence** between the brackets

*While not all simulations have achieved perfect realism, such as Computer Generated Images (CGI) in movies or via recent algorithms such as GANs or diffusion model, which often include distortions or lack certain key details distinguishing them from real images (McDonnell & Breidt, 2010), it is fair to assume that these technical limitations will become negligible in the near future.*

3) on p. 7, "Based on the affective reality theory Makowski (2023)," should probably be "Based on the affective reality theory (Makowski 2023),". In any case, since this construct is interesting and very relevant for the paper, the authors might consider to expand on presenting it a bit further, perhaps with the aid of some clarificatory example.

To-do:

1. **Expand** the **affective reality theory**

Based on the ***affective reality theory*** (Makowski, 2023), which suggests that salient and emotional….

2. Fix typo

4) on p. 8, the authors write "this study does investigate the discriminative accuracy between "true" photos and "true" artificially-generated images (which we consider more a technological issue than a psychological one)". I am afraid I do not understand what is meant by "technological one VS psychological one". I suggest to clarify.

To-do:

1. **Clarify** technological vs psychological

5) on p. 21, when the authors present the "the potential bias induced by face familiarity (as compared to judging completely new items) cannot be discarded", I suggest to specify that the "face familiarity" they are talking about here is that induced WITHIN the experimental setting by re-presenting the same facial stimulus twice, in order to avoid potential confusion with the facial familiarity ratings (and normative ratings).

To-do:

1. **Specify** face **familiarity** as **different** from **facial familiarity ratings**

Although the effect of re-exposure delay was negligible, the potential bias induced by face familiarity**, *that is by the re-presenting the same face stimuli twice*** as compared to judging completely new items, cannot be discarded. Future studies could examine that by incorporating novel face images or increasing the duration of the re-exposure delay