**Revision**

We would like to first apologize for the delay in submitting this revision, in what was professionally and personally an eventful year for some of the authors. We extend our deepest appreciation to the reviewers and the editor for their time, insightful feedback, and encouragement throughout the review process. We are grateful for their recognition of the relevance and interest of our project. We have addressed and responded to all comments and suggestions, with changes highlighted in bold in the manuscript.

**Editor and Reviewer comments:**

I have received three thoughtful reviews of the manuscript that you submitted to Acta Psychological (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.

We appreciate the reviewer's positive feedback regarding the relevance and the theoretical and applied justifications of our study. We agree that the connection between attractiveness and trustworthiness/familiarity should be expanded for readers who may not be familiar with the literature. We added the following clarifications:

Lines 120-122:

Additionally, other factors, such as the stimuli’s perceived self-relevance (Goldstein, 2009; Sperduti et al., 2016), as well as familiarity (Begg et al.,1992), could also play a role in guiding our appraisal of a stimulus. **For instance, Miller et al., (2023) reported that participants were more likely to mistakenly identify AI-generated faces as real because they perceived them as more familiar.**

Lines 132-139:

“Due to their popularity as a target of CGI technology and the prospect offered with facial features that can be experimentally manipulated, AI-generated images of faces are increasingly used to study face processing (Dawel et al., 2021), in particular in relationship with saliency or emotions, as well as to other important components of face evaluation, such as trustworthiness or attractiveness (Balas & Pacella, 2017; Calbi et al., 2017; Sobieraj & Krämer, 2014; Tsikandilakis et al., 2019). [...]. **Similarly, when participants are informed that faces are AI-generated, the perceived artificiality leads to lower trust ratings (Wang & Nishida, 2024), even when they are real faces (Liefooghe et al., 2022). In contrast, when participants are unaware that the faces are AI-generated, trust ratings for these synthetic faces tend to increase (Nightingale & Farid, 2021).** Whereas this line of evidence suggests that reality beliefs have an effect on face attractiveness and **trustworthiness** ratings, the opposite question - whether attractiveness and **trustworthiness contribute** to the formation of reality beliefs has received little attention to date.”

**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.

We agree with the reviewer that additional information regarding the stimuli would be helpful in assessing its suitability for our project, hence the following changes were made (lines 211-214):

“The AMFD (J. M. Chen et al., 2021) is a recently validated database including a set of 110 pictures of homogeneous quality featuring diverse faces (particularly in terms of ethnicity), each (except one) posing with either a neutral or smiling expression. We selected all 109 neutral images (89 women and 20 men) to reduce the influence of confounding factors like affect.”

Additionally, while we did not directly control for certain dimensional properties, such as dominance, competence, warmth, affective valence of expression, and racial ambiguity, we included all the available stimuli - which the authors selected for their variety - to hopefully mitigate the role of these features. We added the following point in the discussion:

Lines 529-532:

“Finally, it is important to note that although consistent in their directions across models and variables, the magnitude of the effects found in the study was relatively small, suggesting that the facial appraisals measured in the study were not the key determinants of simulation monitoring. Hence, beyond exploring new potential mechanisms, future studies should include a more thorough debriefing to try to capture what conscious strategies (if any) the participants used (e.g., focusing on some features of the stimulus - like hair or eyes in the case of faces) to guide their reality beliefs. **Additionally, the role of specific facial features, like perceived dominance, warmth or gender, would be an interesting avenue to explore in future studies, in particular with paradigms directly manipulating these dimensions (for instance using AI to generate faces of different characteristics).**”

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?

We acknowledge the reviewer's concern regarding the lack of justification for the scales used in our study. The inclusion of these measures was primarily guided by expert recommendations (in particular from forensic psychologists) during the study’s design phase. These discussions highlighted the role of traits like narcissism in deception detection, supported by research showing their potential influence on reality judgments (Turi et al., 2022). To address this, we have added the following details to the manuscript:

Lines 100-104:

“Furthermore, dispositional traits, such as high levels of narcissism and low levels of openness and conscientiousness, have been associated with greater susceptibility to fake news (Piksa et al., 2022; Sindermann et al., 2020). **Interestingly, a recent review suggested that narcissism was related to a strong self-perceived ability at detecting lies (Turi et al., 2022), which could translate to participants scoring high on narcissism providing more clear cut and confidence responses. Conversely, those high in honesty-humility tend to be more conservative in their judgments to ensure fairness (Liu et al., 2020), likely resulting in lower confidence ratings.”**

Furthermore, we added the following methodological information (Lines 207-210):

**“**The AMFD is a recently validated database including a set of 110 pictures of homogeneous quality featuring diverse faces (particularly in terms of ethnicity), each (except one) posing with either a neutral or smiling expression. We selected all 109 neutral images (89 women and 20 men) to reduce the influence of confounding factors like affect. **The decision to present the faces for 500 ms was based on pilot studies, which demonstrated that this duration provides a sufficient perceptual window for decision-making and aligns with previous research indicating stable judgment levels and increased confidence beyond this exposure time (Willis & Todorov, 2006).”**

The use of these two scales reflects our belief that attractiveness and beauty, while related, might represent distinct constructs. By employing both scales, we aimed at capturing more nuanced dimensions: the subjective and personal appeal of attractiveness and a more “objectively” judged notion of beauty. We added the following (Lines 227-232):

“Notably, as facial attractiveness is a multidimensional construct, encompassing evolutionary, sociocultural, biological as well as cognitive aspects (Han et al., 2018; Rhodes et al., 2006), we assessed attractiveness using visual analog scales, measuring general Attractiveness (“I find this person attractive”) and physical Beauty (“This face is good-looking”). **This dual-scale approach aims to reflect two conceptually distinct dimensions: Attractiveness might capture personal, Self-relevant and subjective appeal, whereas Beauty might be related to a more “objective” decision based on aesthetic criteria that can be recognized independently of personal attraction. In other words, we wanted the experiment to be able to potentially capture scenarios where a face could be judged beautiful yet not, attractive and vice versa.”**

**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:

We would like to thank the reviewer for acknowledging the significance and increasing relevance of the topic addressed in our paper. We also appreciate the positive feedback on our research design and the careful selection of measures.

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.

In response to reviewer 2 and 3 suggestions, we have reviewed the highly relevant paper mentioned and incorporated its findings into our manuscript (lines 120-122).

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.

While many of our hypotheses are indeed exploratory, we agree with the reviewer that additional rationale and justification are needed.

In response, we have elaborated on the key factors of interest, incorporating feedback from reviewer #1. Specifically, we have provided a clearer explanation of how familiarity (Miller et al., 2023) and trustworthiness (Wang & Nishida, 2024; Liefooghe et al., 2022; Nightingale & Farid, 2021) may influence the formation of reality beliefs. Please see lines 120-139 for these clarifications, and refer to our response to reviewer #1.

We have also expanded on the Affective Reality Theory in response to reviewer #3. Please refer to lines 144-162 for this elaboration.   
  
Additionally, we have also clarified the potential of inter-individual characteristics, such as narcissism and honesty-humility, influence on reality judgments, as well as attitudes towards AI. These clarifications have been made to the following lines 100-104 and can be seen in our response to reviewer #1.

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?

We added the following information (lines 215-220):

The AMFD (J. M. Chen et al., 2021) is a recently validated database including a set of 110 pictures of homogeneous quality featuring diverse faces (particularly in terms of ethnicity), each (except one) posing with either a neutral or smiling expression. We selected all 109 neutral images (89 women and 20 men) to reduce the influence of confounding factors like affect. **The AMFD primarily features racially ambiguous faces, representing multiple racial categories such as multiracial, Latinx, and white. The database includes 81 faces from individuals self-reporting two racial backgrounds and 29 from those with three or more racial backgrounds: 33% Asian/White, 22% Latinx/White, 11% Asian/Latinx, 6% White/Middle Eastern, 5% Black/White, and 5% Asian/Middle Eastern, with about 18% identifying as other racial backgrounds.**

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)?

We clarified as follows (lines 207-210):  
  
In the second part of this study, images of neutral-expression faces from the American Multiracial Face Database (AMFD, J. M. Chen et al., 2021) were presented to the participants for 500ms each, in a randomized order, following a fixation cross display (750 ms). **The decision to present the faces for 500 ms was based on pilot studies, which demonstrated that this duration provides a sufficient perceptual window for decision-making and aligns with previous research indicating stable judgment levels and increased confidence beyond this exposure time (Willis & Todorov, 2006).**

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

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

All the analyses with the questionnaires can be considered between-participants (i.e., each participant has one score on a given dimension, rather than repeated measures as is the case with the main task). However, we would like to underline that the distinction between within & between analyses is not that clear cut in this study as in more traditional paradigms, as the between-participants variables were actually included within the mixed models that contained the within-participants variables (in other words, both analyses levels are carried out jointly - which is a benefit from a statistical perspective).

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.

Missing data was prevented by employing a forced-choice questionnaire, requiring participants to complete all survey items before advancing to the next section, thereby eliminating the need for data imputation.

We clarified the participant exclusion process (lines 252- 256):

“We excluded 5 participants that either failed 2 (>= 66.6%) or more attention check questions, took an implausibly short time to finish the questionnaires or had incomplete responses. **Out of the 5 participants excluded, 2 participants were excluded because they failed 2 out of 3 attention checks, 1 because they did not answer the sexual orientation question, which made further analysis impossible, and 2 had an abnormally low agreement (r < 0.1) between the beauty and the attractiveness ratings (possibly indicating random responses as these two scales exhibited a higher correlation for the other participants).”**

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?

Participants who did not provide sexual orientation were excluded from the analysis (n=1), and for those attracted to both genders, both stimuli genders (i.e., all stimuli) were marked as “relevant” for analysis (note that all participants saw all stimuli, but it is at the analysis step that we focused on the sexually-aligned images).

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

Cronbach's alpha has now been calculated for all scales used in the study. Due to the large number of scales, we will provide the raw alpha values here and update the manuscript to highlight only the lower alphas, directing readers to the supplementary material for the complete set of alpha values. The following changes have been made to lines 199-203:

“In the first part of the study, participants answered a series of personality questionnaires presented in the order below. These include the Mini-IPIP6 (24 items, Sibley et al., 2011) measuring 6 personality traits, the SIAS-6 and the SPS-6 (6 items each, Peters et al., 2012) assessing social anxiety levels, 5 items we devised pertaining to expectations about AI-generated image technology (“I think current Artificial Intelligence algorithms can generate very realistic images”), of which we mixed with items from the general attitudes towards AI scale to lower the former’s saliency and the possibility of it priming the subjects about the task, (GAAIS, Schepman & Rodway, 2020) the FFNI-BF (30 items, Jauk et al., 2022) measuring 9 facets of narcissism; the R-GPTS (18 items, Freeman et al., 2021) measuring 2 dimensions related to paranoid thinking; and the IUS-12 (12 items, Carleton et al., 2007) measuring intolerance to uncertainty. Self-rated attractiveness was also assessed using 2 items - one measuring general attractiveness (“How attractive would you say you are?” Marcinkowska et al., 2021) and the other measuring physical attractiveness (“How would you rate your own physical attractiveness relative to the average,” Spielmann et al., 2020). 3 attention check questions were also embedded in the surveys. **All Cronbach's alpha values were within the acceptable to excellent range, except for the neuroticism subscale of the Mini-IPIP6 and the negative subscale of the GAAI, which were poor, and the Expectations about AI scale, which was questionable (Gliem & Gliem, 2003; see supplementary material for details of the reliability analysis).”**

The rest of the alpha’s are:

Mini-IPIP6   
 - Extraversion (𝞪 = .82)  
 - Conscientiousness (𝞪 = .73)  
 - Openness (𝞪 = .73)  
 - Honest-Humility (𝞪 = .70)  
 - Agreeableness (𝞪 = .82)  
SIAS-6  
 - Social Interaction (𝞪 = .84)  
 - Social Phobia (𝞪 = .87)  
GAAIS  
 - Positive (𝞪 = .84)  
FFNI-BF  
 - Agentic narcissism (𝞪 = .78)  
 - Antagonistic narcissism (𝞪 = .83)  
 - Neurotic Narcissism (𝞪 = .79)  
R-GPTS  
 - Reference (𝞪 = .85)  
 - Persecution [I think](𝞪 =.93)  
IUS-12  
 - Prospective Anxiety (𝞪 = .80)  
 - Inhibitory Anxiety (𝞪 = .82)  
Self-Rated Attractiveness (𝞪 = .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>

We appreciate the reviewer's acknowledgement of our commitment to open science practices and the added strength it brings to our paper. We also thank the reviewer for the guidance on creating a view-only link. As requested, we have provided the link here: https://osf.io/e4pxh/?view\_only=803ae57f0ada4cc5b914081ba2bf567a

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.

We have added the following deviations from pre-registration to the manuscript in a footnote:

“This approach diverges from the preregistration in several key ways. First, the phrasing of items was modified from "Assuming the face you saw was of a real individual, how..." to "I find this person..." A new attractiveness scale (i.e., Beauty) was introduced to capture a more objective measure of attraction. Finally, the data analysis method was altered from Bayesian Mixed Models due to computational limitations, as we were unable to run these models on a high-performance cluster.”

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)?

Our study employed a fully crossed design and therefore no variables were nested in our analysis. In terms of clarifying the result section, we have made the following changes:

Lines 266-267:

“The models included the participants and stimuli as random **intercepts with no nested variables**.”

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.

We decided to run each key appraisal variable in separate models as they were theoretically independent *outcome* variables. However, some of them were included as predictors as well in other models (e.g., familiarity) to adjust for its potentially confounding factors. Attractiveness and beauty, in particular, were kept separate, as they were expected to be correlated, and adding them to the same model would lead to collinearity issues. It would have been interesting in theory to indeed understand the interaction effects of the various appraisals, but on top of the collinearity issues mentioned above, having a 4-5-way interaction would become difficult to interpret. Delineating these dimensions should probably be done in a more thoughtful way experimentally by orthogonally manipulating them.

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?

We have directly tested the effect of exposure delay (which is a function of presentation order). Its effect was quite negligible so we did not adjust for them in the models to not overload these with models with parameters.

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.

We agree that this is a complex study articulating - in a fairly exploratory fashion - quite a number of different concepts. However, we feel like discarding some of it would give a distorted view of the study, hence the reason to be as comprehensive as possible. We hope that our work in response to this round of review helped clarify the manuscript.

**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 complementary 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"

We rephrased accordingly (lines 498-502):

“Another issue is the impact on reality judgements of the prior explicit instruction that “about half of the faces were AI-generated and the other half real photos”. Given this prior information given to participants, it might seem like our enthusiasm pertaining to the finding that most people did indeed believe a high number of stimuli to be fake might be unwarranted, since it simply affirms participants followed the instructions. However, even if that was the case, the finding that our beliefs of reality can be so easily re-programmed with simple instructions and lead to high-confidence answers remains an interesting phenomenon. **Moreover, it is to note that the paradigm did not explicitly instruct participants to balance their answers according to a certain distribution (e.g., 50-50) - merely providing them a description of the dataset (but participants could, and in some cases did, deviate substantially from the information provided)**.”

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)

We have now reviewed and included this study in lines 120-122.

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!

We understand the reviewer's concern about our statement regarding the creation of near-perfect simulations being unprecedented in human history. We agree that this characterization may not fully capture the historical context of photographic manipulations. Consequently, we have revised the manuscript to better reflect the point that current technologies, while not the first to simulate reality, represent a significant shift due to their accessibility, speed, and affordability (lines 51-53):

“Advancements in technology have now made it possible to create near-perfect simulations that are indistinguishable from reality with an ease, affordability and accessibility that are unprecedented 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.

We changed it as follows (lines 64-68):

“While not all simulations have achieved perfect realism, such as Computer Generated Images (CGI) in movies or algorithms such as GANs, which often include distortions or lack certain key details distinguishing them from real images (McDonnell & Breidt, 2010), it is reasonable 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.

We appreciate the reviewer’s comments on the relevance and intriguing aspects of affective reality theory in relation to our study. We have expanded the discussion of this theory in lines 144-161:

“AI-generated content, in particular realistic images, is becoming commonplace and carries important risks for misinformation and black-mailing (Viola & Voto, 2023), emphasizing the need to understand the different components that come into play in the formation of reality beliefs. This exploratory study primarily aims at investigating the effect of facial attractiveness on simulation monitoring, i.e., on the beliefs that an image is real or artificially generated. **The Affective Reality Theory (Makowski, 2023; Makowski, 2018) posits that the default tendency is to believe that experiences are real, with emotional and bodily reactions playing a pivotal role in reinforcing or challenging this belief. According to the theory, there is a quadratic (inverse U-shaped) relationship between affect and reality judgments: stimuli that elicit mild to moderate emotional and/or bodily reactions tend to enhance the perception of realness, increasing confidence in reality. However, when emotional or bodily responses become too intense or overwhelming, the default belief shifts towards non-reality beliefs as an emotion regulation mechanism. In other words, extreme emotional reactions can potentially trigger beliefs of reality denial (“it cannot be real”) as a protective mechanism to help individuals cope with distress. The present experiment can be put in relation with the first part of this proposal - that of a positive relationship between embodied or emotional reactions and appraisals of reality. Following this hypothesis, faces rated as either highly attractive or unattractive - and eliciting stronger reactions - would likely be judged as real. We expect a similar pattern with trustworthiness, where faces judged as highly trustworthy or untrustworthy will be more likely to be perceived as real. Finally, we anticipate a positive relationship between familiarity and perceived realness, as familiar faces tend to be judged as more real (Miller et al., 2023).**”

Additionally, the noted typo has been corrected.

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.

We added the following (lines 166-169):

“Additionally, we will further explore the link shared by dispositional traits, such as personality and attitude towards AI, with simulation monitoring tendencies. Importantly, this study does investigate the discriminative accuracy between "true" photos and "true" AI images (which we consider more a technological issue than a psychological one), focusing on the beliefs that a stimulus is real or fake, independently of its true. **In other words, the present study investigates the psychological process that leads to different *beliefs* of reality, rather than of the discrimination between real faces and actual AI-generated ones, which largely depends on the technological quality of the AI-generation process.**”

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).

The following changes have been made to the manuscript addressing the reviewers suggestion:

Lines 488-489:

“Several limitations have to be noted. The current experimental paradigm required participants to judge the realness of faces they had prior exposure to (which was done to prevent reality judgements from influencing the other ratings). Although the effect of re-exposure delay was negligible, the potential bias induced by face familiarity, **that is by re-presenting the same face stimuli twice**, as compared to judging completely new items, cannot be discarded.”