REVISION

Once again, we would like to thank the editor and the reviewers for their patience and for the encouraging feedback on our manuscript. We have tried to answer and address all the concerns raised to the best of our abilities.

# Reviewer #1

I would like to thank the authors for addressing my comments. In my opinion, this is a nice first study into an interesting area and idea, which warrants publication and follow-up research. It is also an involved study: getting participants to a lab for two sessions, connecting them to several pieces of equipment, and all that during a time with COVID restrictions. I also would like to thank you for your openness and transparency regarding the sub-optimal sample size and lack of preregistration.

While I believe this manuscript warrant publication, there are some aspects that could improve it, mainly making it easier for the reader to follow.

MAJOR suggestions:

I would consider reframing the results around your outcome variables, rather than the two predictor variables (ToM and interoception). In the end, you want to explain the outcome variables, so it makes sense to discuss them as a paragraph on confidence in giving a convincing response, reaction times, heart rate, and potentially dispositional lying. It also aligns more with the figures. I think it would make it easier to see, at a glance, that e.g., confidence is predicted by several indicators, while heart rate is not.

We thank the reviewer for the suggestion to improve the clarify of the results section. As our hypotheses are anchored in the contribution of *ToM* and *interception* abilities on deception skills, we believe that discussing each mechanism separately would maintain the originally intended focus of this paper and be in line with the introduction.

Nevertheless, we do agree that the clarity of the results section can be further improved. We will restructure this section according to your suggestion below.

\*\*\* The discussion starts with "[…] when participants were presented with (fake) physiological feedback (the polygraph condition), instead of a face of a person they had to lie to (the interrogation condition), their response time for both lies and truths increased, as did their heart rate." Yet, for heart rate, the results state: "The heart rate was significantly more elevated during lies as compared to truth […], and during interrogation as compared to the polygraph condition […]." So, this seems contradictory: did heart rate increase in the polygraph condition (as the discussion suggests) or in the interrogation condition (as the results suggest)? Was there a difference between lies and truths (as the results suggest) or not (as the discussion suggests)?

Blabla we acknowledge our mistake and removed the corresponding discussion.

Interestingly, it seems to us like the actual pattern of result (interrogation condition leading to faster responses and higher heart rate) would suggest a heightened state of arousal in the interrogation condition. *TAM elaborate?*

We acknowledge the post-hoc interpretation by phrasing it as such:

*BLABLA but in a tempered way.*

MINOR suggestions:

\*\* (If the MAJOR change to the results is not made) In the results, ensure that each paragraph consistently covers all variables of interest. Now, some paragraphs do state that there were no associations with heart rate, reaction times, and/or dispositional lying traits, but other paragraphs do not specify one of those variables. I can only assume that it means they were not significant either, but it might be good to consistently mention all variables.

Thank you for providing an alternative suggestion. We have worked on the results section to (1) improve the consistency in the presentation of the data and (2) ensure that all associations, significant or not, were mentioned.

\*\* A bit later in the discussion, p. 19, lines 46-51, it is stated that the study adds to the "discourse surrounding the use of physiological measures in past deception research, further questioning its validity as an indicator of deception". Yet, what comes before this sentence suggests that physiological measures (suggestive of a polygraph) reduce lying; it's not so much about the validity as an indicator of deception. This also seems to contradict a later point that the present study was not about objective lying.

TODO

Indeed, this sentence was a bit too strong and overgeneralization, thus we changed it to the following:

“Although physiological indices might yield different patterns between…”

\*\* In the second part of the discussion, different references for the mixed findings regarding interoception and ToM are presented on pp. 19-20, lines 60-4 and p. 20, lines 19-21. Is there a reason for this? Could the repetition of the statement be avoided, and all necessary references be included in one citation?

TODO

\*\* In the methods, when describing the Yoni task, could you clarify the difference in trials assessing affective and cognitive ToM abilities? Perhaps an example is enough.

We have added examples to illustrate the difference between the cognitive ToM and affective ToM trials. In lines XX to XX:

“More specifically, in first-order trials, participants were instructed to make inferences about Yoni's mental state with regards to the objects surrounding it (e.g., "Yoni is thinking of..." for cognitive ToM trials or "Yoni likes..." for affective ToM trials). In more complex second-order trials, participants had to correctly infer the interaction between Yoni and others' mental states (e.g., "Yoni is thinking of the fruit that ... wants" for cognitive ToM trials or "Yoni likes the fruit that ... likes" for affective ToM trials).”

\*\* A few notes about Figure 1:

\* Would a similar figure regarding reaction times be useful?

The figure being already quite “heavy” and complex, we are worried it might add too much weight to the manuscript to add another similar figure. Since the confidence results are the more robust and clear, we believe it is preferential to keep it as the sole illustration – though we would be happy to add a similar one for RTs if the reviewer or editor thinks it will clarify rather than blur the paper’s main take away points.

\* The description "confidence that one tells a convincing lie" or "lying confidence" does not hold for the truths on the right-hand side. Why not something like "confidence that one gives a convincing response"?

We have changed the figure accordingly.

\* The results state "The Focus interoception score was significantly associated with an increased confidence in truthful responses in both the polygraph", but I don't see this asterisk in the figure (only the blue one for interrogation).

We thank the reviewer for pointing out this discrepancy. While revising this manuscript, especially for the results section, we adopted a stricter standard for our significance index. Therefore, the association between Focus interoception score and confidence in truths in polygraph condition–previously “marginally” significant (β = 0.15, 95% CI [−0.02, 0.32], pd = 95.76%) - is now interpreted as non-significant. We have revised the figure to reflect the revised results accordingly.

There are some minor mistakes in language:

\* When comparing 'lies' (plural), ensure to also say 'truths' (plural). Especially the methods and results often discuss 'truth' and 'lies'.

\* P. 4, line 55: remove "a metacognitive dimension of interoception"; already explained above

\* P. 5, line 26: "further reported that individuals"

\* P. 7, line 9: the phrasal verb is "pertain to", but perhaps you mean "containing the item Yoni is referring to"?

\* P. 8, lines 10-33: This technically is one very long sentence. Perhaps it can be split into: "Given interoception's multidimensional nature, the MAIA-2 (Mehling et al., 2012), a 37-item questionnaire, using a 5-point Likert scale, was also administered. It contains eight distinct facets including […]."

\* P. 8, lines 38-41: Perhaps add a note like "taken from their previously completed Autobiographical Memory Questionnaire […]"

\* P. 10, lines 12: Abbreviations ECG and EDA are first mentioned, but not written out.

\* P. 10, line 58: AMQ has been written out, so can just use the abbreviation

\* P. 13, line 4: "The study was not pre-registered (stemming out from […]" (remove 'out' after stemming)

\* P. 20, line 39: "al. (2022) further reported that"

\* P. 20, line 44: "Mohr et al. (2023) found that"

\* P. 22, line 21: "lying skills - rather than of deception self-confidence"

\* P. 22, line 39: "concerns regarding its validity have been"

\* P. 23, line 14-16: "could emerge with sufficient statistical power"

These have been addressed.

Ensure to follow APA rules regarding numbers in the text: use numerals to express numbers 10 or above and write out numbers as words to express numbers up to nine. Don't start a sentence with numerals. Specifically:

\* p. 5, line 42: two conditions

\* p. 6, line 18: Thirty university

\* p. 6, line 57: four colored

\* p. 8, line 12: eight distinct

\* p. 11, line 10: three cognitive-behavioural

\* p. 11, line 20: three electrodes

\* p. 12, line 29: three outcome

\* p. 13, line 53: three Yoni-task … two BES

\* p. 14, line 3: eight MAIA … three HCT

These have been addressed.

# Reviewer #2

I thank the Reviewer for incorporating my suggestions and comments in this revised version of the manuscript which I think is very much improved. I do not have additional comments but I would like to better taken into account a couple of my prior remarks.

The authors cited great work on EF and interoception and lying. They could also think to cite another recent article in which the authors tested different EF and their link with lying: Battista, F., Otgaar, H., Mangiulli, I., & Curci, A. (2021). The role of executive functions in the effects of lying on memory. Acta Psychologica, 215, 103295.

We thank the reviewer for the suggestion. We have included the work by Battista et al. (2021) in our discussion. On line XX-XX:

“Another possibility that should be tested in the future is that of a mediating role of executive functions, given their association with lying (e.g., Battista et al., 2021; Debey et al., 2012) interoception (Molnar-Szakacs & Uddin, 2022).”

2.

Did the authors carried out an a priori analysis to determine their sample size? Please, report this information and if so, please specify on which parameters they based their power analysis. If not, I think the authors should include a sensitivity/posteriori analysis.

This study was not preregistered, and no power analysis was performed to determine the sample

size (mostly due to time available and other constraints related to this being part of a student's

final year project). To compensate for these major flaws, we have taken a variety of steps,

including (in our opinion) an appropriate and conservative statistical treatment (with effect

uncertainty quantification and report), a careful discussion emphasizing the limitations, and most

importantly, a complete transparency and reproducibility.

On a side and tangential note, we have ourselves carefully examined the data (acting as the first

skeptics) and only because we are confident these are interesting patterns did we submit them for

publication. Naturally, collecting more data would have been the best, although impossible due to

the aforementioned reasons, so treating this study as a preliminary proof-of-concept paper

presenting the paradigm and some leads to further investigate and confirm seemed to us like the

best option.

We have also investigated the sensitivity/posteriori analysis mentioned by the reviewer, but from

our research in seems mostly in the case of discrete outcomes where one can estimate the

sensitivity of their predictive classification models. We would be very interested in any pointers

for this type of analysis that the reviewer might have.

Please, add a footnote where you explain this.

TODO

3.

Also, I was surprised to see the authors used physiological measures to understand deception but

I understand that the authors mainly used physiological measures to assess interoception and also

as an additional measure to self-report ones. However, in their discussion they also claim that

physiological measures are good lie detection cues. But, there is a large amount of studies

showing that actually physiological measures are not good indicators of deception as they are

influence by several individual and situational factors. I think this literature needs to be take into

consideration at least while discussing the achieved results.

We agree with the concerns pertaining the accuracy of physiological measures in assessing

deception ability. We have highlighted the debate surrounding its implementation as a valid

measure in the current body of literature and added the following clarifications (l. 333-336):

"By extension, our study adds to the controversial discourse surrounding the use of physiological

measures in past deception research, further questioning its validity as an indicator of deception

(Oviatt et al., 2018; Rosky, 2013)."

Here, I still think the authors should better frame the debate being more specific on the critiques moved to physiological measurements.

TODO

This comment lines up with the point from reviewer 1. We tempered down the sentence to avoid overgeneralization and extrapolation: TODO

Be more specific about the fact that we used only Heart rate and so don’t generalize to “physiological signals” in general.

4. Finally, a very general comment. Could it be that the concepts of interoception and TOM work together, that is could it be that they influence the ability to lie simultaneously and not in a

separate way as it seems was conceived by the authors?

We thank the reviewer for the suggestion, and in essence we very much agree. Indeed, there has

been some research linking interoception and ToM, as well as their neurophysiological

underpinnings. However, much of this research seems focused on emotion processing, which only

constitutes one of the host of cognitive processes required to engage in deceptive behaviour (see

Shah, P., Catmur, C., & Bird, G. (2017). From heart to mind: Linking interoception, emotion, and

theory of mind. Cortex; a journal devoted to the study of the nervous system and behavior, 93,

220-223.). Furthermore, given the overlaps in the literature surrounding ToM and empathy, it

remains unclear whether interoception works with ToM or empathy (specifically affectivce

empathy) in the processing of emotions. As such, while we do not reject the possibility of

interoception working with ToM in influencing lie ability, considering the current gaps in

literature, separating the two constructs and their underlying constructs appeared to be a useful

first approach to delineate potential "main effects" of these processes. However, future studies

(with a different design and a larger sample) could investigate the interaction (and possible

mediation effects) between interoception and ToM by means of, for instance, structural equation

modelling.

I thank the authors for their explanation. I suggest them to add this speculation also in their paper, specifically in the Discussion.

We have incorporated this discussion in our limitation section, on lines XX-XX:

“Finally, there has been some research in the extant literature linking individual differences in ToM and interoception, as well as their neurophysiological underpinnings (Shah et al., 2017; Gao et al., 2019; Ondobaka et al., 2017). As such, it remains a possibility that the two constructs interact in influencing lying ability. However, much of this research seems focused on emotion processing, which only constitutes one of the hosts of cognitive processes required to engage in deceptive behaviour (e.g., Shah et al., 2017). Furthermore, given the overlaps in the literature surrounding ToM and empathy, it remains unclear whether interoception works with ToM or empathy (specifically affectivce empathy) in the processing of emotions. Considering the current gaps in literature, the present study investigates the influence of individual differences in ToM and interoception on lying ability separately; this could be a useful first approach to delineate potential "main effects" of these processes. Nevertheless, future studies (with a different design and a larger sample) could investigate the interaction (and possible mediation effects) between interoception and ToM by means of, for instance, structural equation modelling.”