REVISION

We thank the editor and the reviewers for giving us the opportunity of revising and substantially improving our manuscript.

# Reviewer #1

We thank the reviewer for providing these thorough comments that helped us clarify and address several important points in the manuscript and the abstract.

I do think the strenuous efforts in the analyses reduce the concern that findings might be Type I errors. However, these limitations of the study are only addressed at the end of the methods section. I feel that a different title (e.g., "a preliminary investigation into …") and a mention of the small sample could be made in the abstract.

We agree with the reviewer that the abstract could be more detailed and forthright for the sake of transparency. We have made changes to the title of the paper and the relevant portion of the abstract now reads as follows (l. 10-16):

*“The present study (n=26) examined the contribution of Theory of Mind (ToM) and interoception on our ability to lie using a directed lie paradigm with two conditions (“Interrogation” and “Polygraph”), designed to enhance each of the two mechanisms. In view of the relatively small sample size, we performed a feature reduction on the measures used to assess our two groups of predictor variables using factor analysis and extracted a unique factor of ToM as well as 4 components of interoception (Metacognition, Listening, Focus and Regulation).”*

You describe that previous literature has focused on deception detection, but not deceiving ability. I think this is logical, as it may seem more relevant or important to detect lies than to generate them. Could you add an argument why it might (sometimes) be necessary to be able to deceive others?

We thank the reviewer for this suggestion. We have included additional arguments regarding the necessity of the ability to lie as follows (l. 35-41):

*“While often conceptualized to be immoral and unconscionable, lying is ubiquitous in everyday life, and being able to lie skillfully can sometimes facilitate interpersonal relationships, helping us avoid conflict or causing emotional harm to others (E. E. Levine & Lupoli, 2022). In fact, recent research shows that certain forms of deception, such as prosocial lies (i.e., false statements told to benefit others; E. E. Levine and Lupoli (2022)), can increase trust (E. E. Levine & Schweitzer, 2015). Moreover, individuals who told altruistic lies were perceived as more benevolent than those who were honest (E. E. Levine & Schweitzer, 2014).”*

Is interoceptive awareness one of three facets describing "one's ability to identify and appropriately respond to their perceived interoceptive state" (as argued first) or a "metacognitive dimension of interoception"? These do not sound like the same thing to me.

We agree that the two descriptions can be improved for clarity. We have hence amended our initial description of interoceptive awareness (“one’s ability to identify and appropriately respond to their perceived interoceptive state) as follows (l. 60-62):

*“interoceptive awareness (also referred to as interoceptive insight; Khalsa et al. (2018)) - one’s ability to correctly evaluate their interoceptive ability”*

What kind of association between interoception and rational decisions is posited? And are lies risky? Are lies rational or irrational? Could it be that people high in interoception show more immoral behaviour (as detailed in the paragraph below this one), and then choose a rational path of lying because it less risky? Is this literature really contrasting and mixed?

While we recognize that lies could potentially be thought of as rational/less risky depending on the context (such as in laboratory-based economic games where participants are told their behaviours would remain anonymous), to limit the scope of our research, we focused on everyday lies which are commonly construed as risky given the potential reputational loss if they were to get caught (see Kireev, M., Korotkov, A., Medvedeva, N., & Medvedev, S. (2013). Possible role of an error detection mechanism in brain processing of deception: PET-fMRI study. International Journal of Psychophysiology, 90(3), 291-299.). Additionally, we concur that the nature of the relationship between interoception and risky decision making can be better supported and fleshed out in our arguments. As such, we have updated this paragraph as follows (l. 77-87):

*“Indeed, Sugawara et al. (2020) further reports that individuals who received interoceptive training were more likely to show higher interoceptive accuracy and make reasoned decisions. Given that deciding to lie generally involves a consideration of the potential costs of getting caught, and hence could also be perceived as risky behavior (Kireev et al., 2013), interoception could be construed to be negatively related to lying ability. However, some studies have instead found heightened interoceptive attention (one’s self-focus towards internal bodily signals), to predict apathetic, immoral behaviour, such as cheating (Ditto et al., 2006; Lenggenhager et al., 2013; Williams et al., 2016). Extending these findings to social cognition, Vabba et al. (2022) further reports individuals with lower interoception told significantly less egoistic lies when the social reputational stakes were high, whereas individuals with higher interoception did not exhibit a significant difference in the number of lies told.”*

Could you explain (or cite literature as support) that a short RT indicates a better lying ability? It is somewhat logical for higher confidence and lower physiological arousal to indicate a better lying ability, but could it not be that a longer RT means the lie is better formulated and thus reflects a better lying ability too?

Although we agree with the reviewer that a longer RT could be indicative of a better constructed lie, given the experimental context in which participants engaged in lying (participants were asked questions of which the truthful responses had already been recorded), a longer RT would be more suggestive of hesitation to lie. Indeed, this is also consistent with the cognitive load concept outlined in theories of working memory, which posits that lie-telling requires more cognitive resources and hence requires longer RT. We have clarified the relevance of short RT as a proxy for better lying ability as follows (l. 98-102):

*“Consistent with the cognitive load approach outlined in several theories of deception (such as the Four-Factor Theory (Riggio et al., 1987) and Activation-Decision-Construction Model (Walczyk et al., 2014)), as well as previous findings which suggest response time as a reliable cue to deception (Walczyk et al., 2009; Gonzalez-Billandon et al., 2019), we regarded shorter response times as a proxy of better lying ability.”*

What was the rationale for the direct and indirect phrasing of the questions? There seems to be no hypothesis or reason for this, but then it is considered a factor in (some of?) the analyses. What are the results when this factor is not considered?

Given past findings that question phrasing could influence participants’ response time and compliance to task instructions (to lie or tell the truth) (Walczyk & Cockrell, 2022), as well as to simulate interrogation procedures, we modulated the phrasing of the questions to control for potential confounding effects. We have added this clarification as follows (l. 165-168):

“*In addition, as past studies have found associations between lying behaviour and type of question phrasing (Walczyk & Cockrell, 2022), each question was phrased either directly (e.g., “What is your favourite sport?”), or indirectly (e.g., “Is your favourite sport Hockey?”) to reduce possible confounding effects.*”

In the discussion (p. 19, lines 311-313 and further), the validity and reliability of some of the measures is questioned. While I understand that composite scores from the factor reduction are used, it could be good to include information about the self-reported measures (e.g., Lie Scale) in terms of their validity and/or reliability (e.g., Cronbach's alpha).

\*\*\*DOM\*\*\*

There are some language mistakes throughout, for example:

* p. 4, line 43: "Garfinkel et al. (2015)'s conceptualizes" ◊ no apostrophe S, and I guess 'conceptualized' or 'conceptualize' (plural)
* p. 5, lines 70-73: "We expected ToM and interoception … in particular in the Polygraph and the Interrogation condition, respectively." ◊ is it not the other way around, that ToM is linked to the Interrogation condition and interoception to the Polygraph condition?
* p. 5, line 78: "The final sample consists 26 participants" ◊ consists of
* p. 9, line 142: "All participants reported believing that it was indeed their own signals being shown" ◊ that their own signals were shown
* p. 10, line 184: "As all the analyses and data has been made available" ◊ have been
* Abbreviations are inconsistent (TOM vs ToM; Yoni Task vs YONI task) and sometimes used without the full term written out first, so that the reader does not know what the abbreviation stands for (p. 8, line 139: ECG, EDA; explained on p. 9, lines 152-153).

We note the language and formatting mistakes pointed out and have made the corrections accordingly.

# Reviewer #2

We thank the reviewer for the constructive comments as well as the suggestions made to include other works that had prompted us to delve deeper into the extensive literature of deception, interoception and theory of mind, and how their underlying processes may overlap and complement one another.

In their Introduction the authors state that the ability to lie has been always studied in term of lie detection instead of paying attention to the factor that can make people able to lie. This is true, but there are also studies showing that some personality traits are particularly related to deception. For example, people showing traits of the dark tetrad adopt manipulative and deceptive behaviours, especially those showing psychopathic traits. There are quite some studies on this, a very recent one is Eric Rassin, Carmen Sergiou, Dimitri van der Linden & Josanne van Dongen (2023): Psychopathy as a predisposition to lie hedonistically, Psychology, Crime & Law, DOI:10.1080/1068316X.2023.2213802. I think that considering that this literature should be at least cited to provide a comprehensive picture of the state of the art. I think it could also be relevant -on a very general note- for discussing the results achieved.

We recognize that in the attempt to keep our article concise and to the point, we did not include an extensive literature review on the field of deception, specifically in relation to personality traits of individuals with a high propensity to lie. We have included this in the text as follows (l. 32-35):

*“Nevertheless, some findings suggest a relationship between the propensity to tell lies, and traits that characterize the socially malevolent profile known as the Dark Triad (Paulhus & Williams, 2002), such as narcissism (Zvi & Elaad, 2018) and psychopathy (Rassin et al., 2023).”*

On p. 4, the authors mention the link between interoception and individual differences "Interoception has increasingly been tied to subjective perceptual experiences (Connell et al., 2018; Seth et al., 2012), as well as individual differences in executive functions, emotional processing, and decision-making (Barrett & Simmons, 2015; Murphy et al., 2019; Petzschner et al., 2021)." This made me think to the link between lying and executive functions and the possibility that because of the link of EF with interoception and lying (separately) it is reasonable argue a link between interoception and lying.

We agree it was an oversight that we did not draw links pertaining the underlying role of executive functions in interoception and lying. We have added this in the text as follows (l. 65-70):

*“In particular, neuroscientific findings investigating the correlates of interoception have underlined the potential role of the anterior cingulate cortex (ACC) and anterior insula (AI) (Craig, 2009; Critchley et al., 2009; Khalsa et al., 2009; Wang et al., 2019), both of which are often thought to be activated during deception (Abe, 2011; Baumgartner et al., 2013; Sip et al., 2008), and have been implicated in cognitive processes associated with deception (e.g., cognitive control (Molnar-Szakacs & Uddin, 2022), conflict detection (Kerns et al., 2004)).”*

I found a bit unclear the hypothesis. Did the authors have any hps on possible differences between the two manipulations (polygraph and interview)?

We recognize that the phrasing of our hypothesis could be improved. Specifically, we hypothesized a positive link between lying ability (as measured by lie confidence, RT and heart rate) and interoception in the Polygraph condition and a positive association with ToM skills in the Interrogation condition. The part with our hypothesis now reads as (l. 93-98):

*“The Interrogation condition was designed to emphasize (and preferentially mobilize) ToM-related mechanisms, whereas the Polygraph condition was designed to emphasize interoceptive mechanisms. In particular, we expected lying ability (i.e., higher lie confidence, shorter response time and lower physiological arousal), to be positively predicted by individuals’ interoceptive abilities in the Polygraph condition, and by ToM skills in the Interrogation condition.”*

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.  
\*\*\*DOM\*\*\*

I am a bit confused concerning the final sample size. Was it composed of 26 or 25 participants?

We have added further clarification pertaining to the final sample size (l. 107-111).

*“The final sample consists of 26 participants (Mean age = 20.9, SD = 2.0, range: [18, 25], Sex: 65.4% women, 34.6% men). The heart rate of one participant and response time of one participant were excluded from further analysis due to extreme outlying values. To maximize statistical power, the problematic data from these 2 participants were only excluded from analyses involving those measures; all other data were retained for analyses.”*

Was the study pre-registered?

As this study stemmed from an undergraduate final year project, it was not pre-registered. That being the case, the full reproducible analysis script, statistical results report, and data would be made openly available at [**masked for blinding**]. We have included this information as follows (l. 244-246):

*“The analysis was not pre-registered (stemming out from an undergraduate’s final year project), but the full reproducible analysis script, statistical results report, and data, are available at [masked for blinding]”*

I believe more clarification is needed concerning the exact design used in the study. Was it a between design? Which were the exact dependent variables?

We have added further clarification in relation to the experiment design (l. 195-196):

*“A within-subjects design was used in the present study, which is comprised of 2 sessions, to investigate the roles interoception and ToM play in lying ability.”*

We have also rephrased the part of the text concerning outcome measures for clarity (l. 209-213):

*“Three outcome variables were recorded for each trial of the deception task, namely the participants’ confidence ratings that their answers (lies vs. truths) were convincing, the response time (RT) between the question onset and the participant’s key press (indicating the end of their verbal answer), and the change in heart rate associated with the response (within a window of 3.5 s).”*

Similarly, the tools used need much more elaboration. In particular, the Yoni Task and MAIA-2 need to be better explained (scales, some examples of items, etc).

We recognize that the measures used can be better explained. We have included in more details and rephrased significant portions of the text. The relevant texts now reads as follows:

Regarding the Yoni Task (l. 118-129):

*“The Yoni Task (Shamay-Tsoory & Aharon-Peretz, 2007) is a behavioral task which assesses first and second-order ToM abilities in both cognitive and affective domains. Participants were presented with the face of a character named “Yoni”, surrounded by 4 colored pictures of objects or faces, one in each corner of the screen. […]. During each trial, participants were shown a question pertaining the item Yoni is referring to, and asked to make responses based on specific corresponding cues such as the directions of Yoni’s eye gaze, facial expressions etc., In the control trials, participants made judgements based on Yoni’s physical context (physical ToM). 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…”). 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”).”*

Regarding the MAIA-2 scale (l. 145-153):

*“Given its multidimensional nature, the MAIA-2 (Mehling et al., 2012), a 37-item questionnaire which measures 8 distinct facets of interoception including Noticing (****I notice when I am uncomfortable in my body****), Not-Distracting (****I try to ignore pain****), Not-Worrying (****I can stay calm and not worry when I have feelings of discomfort or pain****), Attention Regulation* ***(I can refocus my attention from thinking to sensing my body****), Emotional Awareness (****I notice how my body changes when I am angry****), Self-Regulation (****I can use my breath to reduce tension****), Body Listening (****I listen to information from my body about my emotional state****), and Trust (****I trust my body sensations****), using 5-point Likert scales, was also administered.”*

I was left unclear why the authors included the BES questionnaire. I do not see any hypothesis on empathy and lying, so I wonder why this tool was administered (even if I understand is related to TOM).

We note that empathy and ToM are increasingly regarded as unique constructs that rely on distinct neural mechanisms. However, there are substantial overlaps in the literature relating to their underlying components (cognitive empathy with cognitive ToM; affective empathy with affective ToM). Furthermore, in trying to include a subjective assessment of ToM, we found few validated scales specific to measuring ToM. However, we recognize that the two are discussed as psychologically distinct in emerging research and have hence included the following clarification (l. 391-397):

*“Moreover, although the cognitive and affective components of ToM and empathy share overlaps in the current literature, and there is no consensus regarding how the two concepts should be delineated, recent evidence nonetheless suggests ToM and empathy are necessarily distinct constructs with separable underlying mechanisms (Kanske et al., 2015). As such, future studies are warranted to further investigate the associations between ToM and lying ability using validated instruments sensitive to measuring ToM (such as the Theory of Mind Inventory: Self-Report, Hutchins et al. (2021)).”*

More elaboration is needed also regarding the deception instructions provided to participants. Maybe it could be helpful that the authors add them as a Supplementary Material.

\*\*\*DOM\*\*\*   
  
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).*”

I also wonder whether the authors took into account the possibility that participants while deceiving were actually reporting the truth as I guess the authors did not have a ground truth allowing them to check this. Could this have affected the results?

We thank the reviewer for this suggestion. In fact, the stimuli presented in the deception task were presented as part of the Autobiographical Memory Questionnaire (AMQ) participants were administered in session 1. Since participants were not aware that these items would subsequently be implemented as stimuli, their responses were used to establish the ground truth. We have rephrased the part describing the deception task and AMQ as follows (l. 196-209):

“*During session 1, participants answered a brief demographic survey as well as a questionnaire regarding their personal preferences and subjective experiences (Autobiographical Memory Questionnaire, AMQ), followed by a series of psychological scales […] For all participants, session 2 began with the deception task, followed by the Yoni task and the HCT, with the latter two presented in a randomized order. In the directed-lying task, items in the AMQ were presented as stimuli, with participants’ recorded responses (in session 1) used to establish the ground truth.”*

Also, considering the task proposed I have doubt on people's engagement with lying. Considering the questions were neutral ones, I question how motivated people were to deceive (as I understand they also did not receive any compensation for participating in the study).  
\*\*\*DOM\*\*\*

On p. 8, the authors refer to "behavioural tasks", please clarify in brackets which task are referring to.

We have added the relevant information accordingly (l. 200-202):

“*During session 2, cognitive-behavioural tasks (i.e., the deception task, HCT and the Yoni task) were administered to participants while their physiological signals (including cardiac activity (ECG), respiration (RSP) and electrodermal activity (EDA)) were being recorded.*”

Also, the authors recorded people's reaction times, but did they also have a baseline phase during which they recorded individuals' reactivity?

\*\*\*DOM\*\*\*

The two conditions created remind me the paradigm of challenging memories with social feedback. Did the authors consider such studies? Maybe it could be helpful also for better justify their results.

\*\*\*DOM\*\*\*

I found the discussion of the results a bit hard to be followed. I think the authors should try to clarify how the achieved results fit or are in contrast with previous findings. The did try to but I honestly got lost in the link made and how they explained their results.

We thank the reviewer for the feedback given and recognized that the discussion section could be improved. As such, we rephrased several sections and included more references to relevant studies so as to better explain the results we obtained In particular, we focused on the differences we observed in contrast to past studies

With regards to our findings in relation to ToM (l. 338-348):

“*While previous bodies of work have reported mixed findings regarding the association between interoception and ToM (Chiou & Lee, 2013; Gendolla & Wicklund, 2009; Scaffidi Abbate et al., 2016; Wundrack & Specht, 2023), our results suggest the two are negatively linked. […] future studies are necessary to investigate the interaction of these mechanisms in different social contexts.*”

With regards to our findings in relation to interoception (l. 351-360):

“*Indeed, this is in line with previous studies that found individuals with low interoception were more averse to risk when reputational stakes were high, telling fewer egoistical lies (Vabba et al., 2022). In fact, Vabba et al. (2022) further reports that people with high interoception abilities were less likely to differ in risk-taking tendencies, telling the same number of lies regardless of the social stakes. Consistent with our results, Mohr et al. (2023) further reports that individuals with high interoceptive accuracy were more likely to make egocentric decisions. However, in contrast to previous studies (Füstös et al., 2013; Owens et al., 2018; Pinna & Edwards, 2020; Pollatos et al., 2007), we did not find any significant relationship between individuals’ interoception scores and their heart rate changes during their answers.”*

I think the authors need to do a small effort in clarifying the practical implications of their work.

We have included the practical significance of our work in the text as follows (l. 402-409):

“*To this end, we introduced a new paradigm that is able to modulate the contribution of these mechanisms while remaining relevant to applied fields of lie detection and criminology (in which the experimental conditions find echoing practices). Notably, our results provide princeps evidence that interoception could be a key - and overlooked - mechanism in deception. Furthermore, our findings extend and offer an alternate perspective to the debatable use of polygraphs, suggesting that its utility for lie detection is not only questionable, but could potentially enhance the deceptive ability of people with better interoception.”*

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. 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 appears to be a more appealing approach for the sake of clarity.