Dear Editor and Reviewers,

Thank you for taking the time to review our manuscript titled *A fragmented news environment and the illusion of knowledge*. We sincerely appreciate your valuable feedback and constructive comments, which have helped us improve the quality of the report.
Below, we address each of your comments and provide detailed explanations of the revisions we have made to address the concerns you raised. We hope that our responses adequately address your questions and that the revised manuscript now meets the standards for the next stage.

Sincerely,
The authors

Major changes

We thank the reviewers for their precious feedback; it helped us in improving our research protocol.

When we designed the pre-test survey, we identified the perceived knowledge as a possible confounding variable. Therefore, even though we did not formulate any specific hypothesis regarding perceived knowledge, we decided to use it to classify the topics. The original design aimed to have four groups in order to map all the possible combinations of high/low perceived knowledge and self-involvement.

However, the findings of the pre-test revealed a robust correlation between self-involvement and perceived knowledge (Cronbach's Alpha = 0.87, 0.859 - 0.88 [CI .99]). This posed a challenge in having topics that are highly involving without being perceived as highly knowledgeable, rendering a pure 2x2 design infeasible.

Given that we did not have a specific hypothesis about the role of perceived knowledge, we have now decided to change the protocol to having **three experimental groups, characterized by topics classified as low, medium, and high involving**. As such, our primary focus centers on the variable of interest: self-involvement. We made necessary adjustments to **Table 1** and to **Appendix A** to enhance clarity.

Our choice was furthermore the result of some doubts we had regarding the special prison regime 41bis, after it recently got to the center of public attention, and we have good reasons to doubt its current positioning as a low-involving topic. We had similar concerns about radioactive waste, which was also removed as a possible topic.

Lastly, in order to enhance clarity, we decided to rephrase the title to explain the aim of the study better.

Covariates

[Reviewer 2] I think the rationale behind including these variables is unclear. The authors do not describe any background about these measures, and do not describe any hypotheses for how they would influence the results. I think it should **at the very least be stated explicitly that these are included for exploratory purposes**, unless there are some hypotheses for them. Furthermore, the statement that these variables "will be included as covariates and control variables" is not very specific. Will these variables be included as covariates/controls in all analyses? Or will you first test a model using only experimentally manipulated variables, and later include these as controls? There is no mention of either of these variables in the table on page 8 (PS: table number is lacking here). The role of these variables in analyses should be clearly specified. The current description opens up for analytical flexibility.

[Reviewer 1] it would be valuable to distinguish between Confirmatory and exploratory hypotheses. Second, at a point in the paper, it is said that additional variables with be used as control variables (the ones in appendix d). Then, in the statistical analysis table, they are not used. I would like to have a clear understanding of **how and why they are used, maybe by creating additional hypotheses.**

Authors' response:

We thank the reviewers for raising this point, indeed the presentation of the analyses left room for many degrees of freedom in the analyses. The use of control variables must be intended as exploratory. Even if, in some cases, we can imagine the direction of an effect (i.e., an intuitive cognitive style may enhance the susceptibility to an overestimation of one's knowledge), we will use these additional measures to explore further the obtained results. Following examination, we have also decided to move one exploratory test as a

main prediction, that is the existence of the illusion of knowledge regardless of exposure (new Hypothesis 3, page 10).
We have now clearly distinguished our predictions from the exploratory hypotheses. We included a new paragraph with the list of exploratory analyses (page 12).
Lastly, we have also decided to include a measure of Intellectual Humility at both T1 and T2, for exploratory purposes only.

Data exclusion

[Reviewer 2] No rules for data inclusion/exclusion are described, except for the mention that incomplete submissions will be deleted on page 9. I find the statement about deleting incomplete submissions to be ambiguous. I assume that a response from a participant that for instance failed to answer a single item in the social media use measure would not be deleted – but this is not clear from the manuscript. Again, to prevent analytical flexibility, the authors should **be clear about what "incomplete submissions" mean.** Does it restrict to main dependent variables? Is there a cut-off point (e.g., more than 5% or 10% of responses missing) where a participant will be excluded?
More generally, rules for data exclusion should be described. This also relates to the "control questions" mentioned above: will participants be included if they fail these control questions? Why? Why not?

Authors' response:

Thank you for your constructive suggestions. We have now included a specification of the exclusion criteria in the design protocol (page 14).

Each page of the questionnaire will have a force-answer setting that will not allow participants to skip any questions. Only the submission by participants who abandon the study before completion will be deleted, as dropping out is considered a withdrawal of the consent to participate in the study.

We have furthermore decided to introduce an additional exclusion criterion to control for the misreporting of demographics across the two experimental sessions. Since Prolific provides demographics about participants, it is possible to match this information with the demographics provided by participants themselves at T1 and T2. The case of a mismatch might suggest that someone

is participating by using someone else's account. For this reason, the mismatched submissions will be deleted and thus excluded from the analyses. This criterion will reduce the likelihood that we will receive responses from two different respondents associated with the same participant ID.

We do not plan any other exclusion of data. However, as an exploratory analysis, we will repeat the pre-registered analyses, excluding participants who failed all the control questions (manipulation and attention checks). See the response below regarding the Control questions, and page 8 in the manuscript.

Experimental groups

[Reviewer 1] I also do think that it would be valuable to have **a hypothesis regarding the 4 groups** of thematic in terms of knowledge/involvement (appendix A and Table 1). Table 1 indicates 4 groups but the study design is only about two groups (high emotion and low emotion thematic). Can you clarify the design or Table 1?

[Reviewer 3] the hypotheses and the design are not sufficiently explained or theoretically justified. The proposed experiment incorporates a 2 (Emotional Intensity: High vs. Low) x 2 (Perceived Knowledge: High vs. Low) x 2 (Exposure: Yes vs. no) x Time (t0 vs. t1) mixed design, with the two latter factors manipulated within-subjects. The exposure and time factors are clear (but should be explicitly mentioned in the design section on p. 6).

Authors' response:

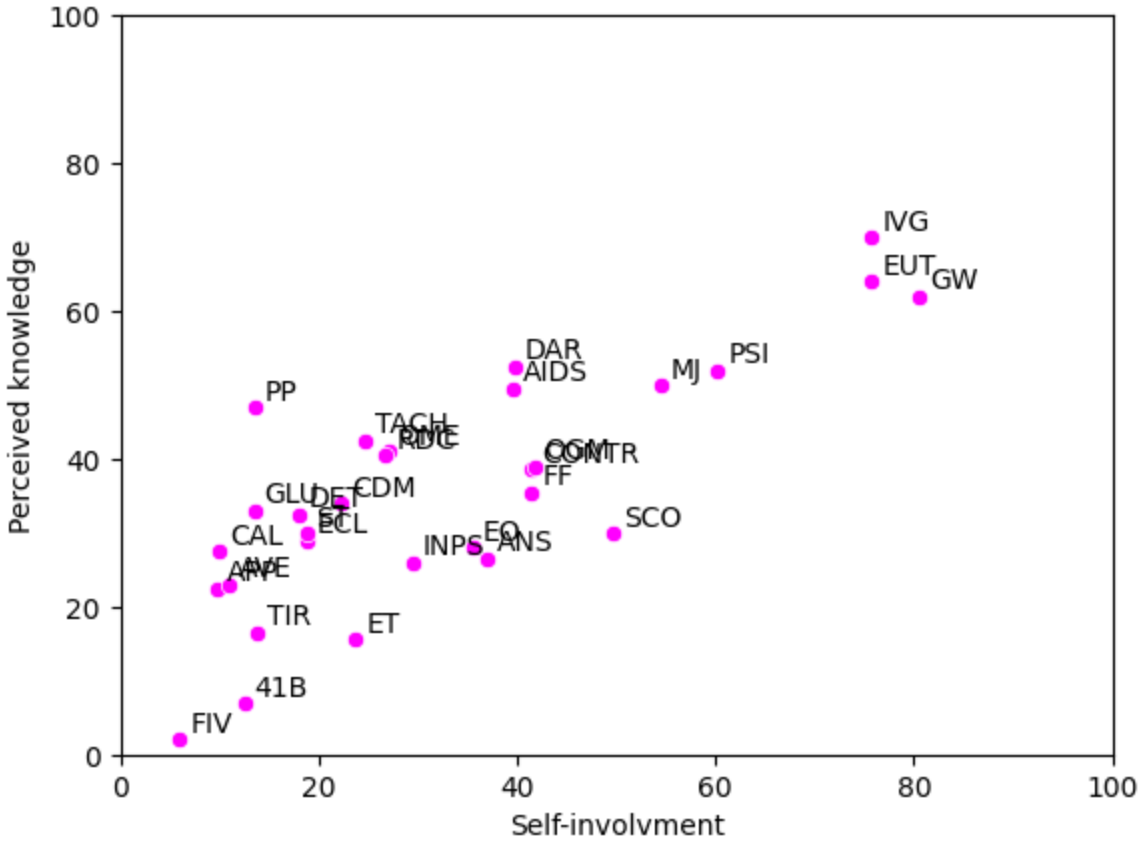
We appreciate the reviewers for highlighting this matter. As stated in the general remark section, we have now reorganized topics under a new categorization that makes more theoretical sense. Please refer to the beginning of this document for the major changes we planned in the experimental groups.

Possible floor effect

[Reviewer 3] I found no explanation in the proposal why topics are a priori selected based on perceived knowledge (high vs. low). Again, I am sorry if I missed something, but why were the materials selected based on high perceived knowledge (baseline) in a pretest? Such an approach is especially prone to floor effects. If participants' perceived knowledge is already very high (e.g., for global warming or abortion in the materials selected from the pretest), then exposure cannot further increase the perceived knowledge

Authors' response:

We thank the reviewer for raising this point. Indeed, the format of the picture presenting the level of perceived knowledge of the several topics may give the impression that the topics with the highest level of perceived knowledge are close to a ceiling. This, however, is a distortion of the graph: for example, the average perceived knowledge of climate change stands at 70/100, leaving ample room for potential shifts due to exposure. Moreover, all knowledge questions have been calibrated to be relatively challenging, mitigating concerns of skewed responses at the lower end. Lastly, our study includes four other topics with varying levels of perceived knowledge and self-involvement (low and medium), ensuring a comprehensive analysis of exposure under a variety of conditions.



The news feed

[Reviewer 1] It is difficult to understand what participants will see in the experiment. Would it be possible to have at least a screenshot of the newsfeed? Are the newspaper titles presented in random order or not? (it is not said through the method section). Can the participants only see the title – text or also a photo? Can they click on it to view the website for each link? We need to know more to improve the possibility to replicate the experiment.

Authors' response:

We thank the reviewer for this crucial information that was missing from the manuscript. We took the opportunity to add this information: the news feed will be composed of news posts (title, image, short description) displayed in random order across participants. Users will be able to react or comment under the news posts but they will not be allowed to open the original articles. We integrated the description of the news feed and a screen of the news feed in the Experimental protocol (page 4).

Data & analysis

Software & Data availability

[Reviewer 1] Finally, it lacks some explanation regarding how the analysis will be performed: with R or another tool – will be the dataset/analysis available on OSF? And if not, why? I wish to emphasize that, as provided in the guideline: PCI RR is a signatory of the Transparency and Openness Promotion (TOP) guidelines, which describe a series of modular standards for transparency and reproducibility in published research. In general, authors are required to make all study data, digital materials, and computer code publicly available (at Stage 2 submission) to the maximum extent permissible by relevant legal or ethical restrictions. While it is a stage 1 manuscript, I would like at least a statement regarding how data, material, and code will be available for stage 2, at best the use of a script on simulated data to understand how you will perform your ANOVAs and t-tests.

Authors' response:

Thank you for highlighting this matter. We have now included this information in the manuscript: we will conduct the analysis using R after having pre-preprocessed the data with Python. All the data, raw and preprocessed, together with the code, will be shared on the OSF repository of the project. We included this specification on page 15.

Statistical tests and hypotheses

[Reviewer 2] the authors plan to use ANOVAs, or a Friedman test as a non-parametric alternative if assumptions are violated. However, to my knowledge a Friedman test cannot test for an interaction in the same way as an ANOVA, and it is thus unclear how the hypotheses proposing an interaction will be analyzed in case of violated assumptions. Perhaps other alternatives such as robust ANOVA could be used instead.

[Reviewer 3] the authors propose a 2 x 2 repeated measures ANOVA for H1-H2. Where do the repeated measures come from if the authors have computed a difference score?

Authors' response:

Based on your kind suggestions, we have made some substantial changes. We will not have the difference score anymore and we formulated the hypothesis

accordingly, given that we now have three distinct groups. All hypotheses will be tested using mixed-effects regressions. We updated the manuscript accordingly (pages 7 - 10).

Effect size and sample

[Reviewer 2] I find the justification of the effect size to lack in detail. The current manuscript refers to an effect size of f = 0.15, stating “the effect size was adjusted based on the results obtained by Schäfer in a similar experimental protocol”. I looked briefly at the findings from Schäfer (2020), and found only one effect size, η2 = 0.01, which converts to a Cohen’s f = 0.10 (using the easystats package in R). So I wonder if I have misunderstood, if the authors are referring to a different effect size, or if something else is going on.
In general, I find this part to lack detail. The authors mention that the sample size is computed over the main and interaction effects, but this should be further explained (the necessary sample size would presumably differ between main and interaction effects).

[Reviewer 3] The proposed sample size of n = 950 is sufficient to detect the proposed effects. However, my own power analysis with the mentioned parameters led to a required N = 580 (interaction/main effect in a 2 x 2 between-subject design). Could the authors add more context to which specific interface in Gpower led to the required N of 768? The sampling plan is transparent.

Authors’ response:

We thank the reviewers for their thorough examination of our power analysis, which indeed was mistakenly modeled after a different test than the one reported. However, following the reorganization of topics into different categories and the change of statistical model used, we have now conducted a new power analysis to estimate the power of our experimental setting under a variety of different effect sizes for H1, H2a, H2b, and H2c. Given the complexity of the statistical model, we have resorted to simulations and to a series of arguably plausible assumptions about the possible values of different variables. Please find the R script used for these simulations attached.

Attrition rate

[Reviewer 2] Another question here concerns the attrition rate. I am not well-versed in studies with a 2-week lag between experimental sessions, but my gut feeling is that 15% is a low estimate of attrition. It would be nice to know whether this expected attrition rate is based on data from similar studies, is a guess, or something else.

Authors' response:

We thank the reviewer for this consideration. We predict a 15% attrition rate based on a similar study conducted by one of the authors on a similar pool of Italian subjects (Ronzani et al., 2022). In said study, data was collected every ten days for several sessions, and the attrition rate between sessions was considerably lower than 15%. We consider the 15% estimate to be conservative, and most likely, we will achieve lower levels of attrition. We have now included the original reference in the text for clarity.

Ronzani, P., Panizza, F., Martini, C., Savadori, L., & Motterlini, M. (2022). *Countering vaccine hesitancy through medical expert endorsement*. Vaccine, 40(32), 4635-4643.

Non-significant findings

[Reviewer 2] The authors make the following statement: "If the test will result non-significant, we cannot rule out that the difference is negligible, that is: there is no difference in the assessment of perceived knowledge of the selected topics before versus after the exposure. If so, it may be that our experiment failed to elicit such an effect, and further analysis will be then required to investigate the results, taking into account other variables."

This is an ambiguous statement. Which further analyses are required? Do the "control variables" come into the picture here? I think the authors should look into whether equivalence testing or Bayesian analysis could be helpful in case of non-significant findings.

[Reviewer 3] "If the test will give non-significant results, we will claim support for the null hypothesis, that is: the emotional intensity does not affect the knowledge illusion." Such a claim is at least problematic for standard frequentist statistics; I suggest Bayesian tests or equivalence tests for this case.

Authors' response:

We thank the reviewers for their kind suggestion. We have decided to conduct equivalence tests when we need to test for equivalence between the groups. We included references to equivalence tests in the Hypothesis section and modified Table 3 on page 15 accordingly.

Emotional intensity

[Reviewer 2] Hypothesis 2 and 4 concern emotional intensity as a moderator. I find these hypotheses to lack clear justification. The only discussion of the background for these hypotheses that I can find is on page 3, paragraph 3, where it is pointed out that previous studies do not control the topics used as stimuli (a good point!), and in the final sentence: “Following Park’s intuition (2001) we believe that the key characteristic that might inflate perceived knowledge is the perceived involvement of the individual, regardless of the topic being assessed: whether it is political, scientific, health-related, and so on.”
This strikes me as **insufficient for proposing the emotional intensity-hypotheses**. It is not clear from these general observations that the effects of exposure should be stronger for emotionally intense topics, and the authors should expand on why they propose hypotheses in this direction. There are studies on related topics, for instance this study (https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.1836?casa_token=5tsruAIICHwAAAAA:OSeAXIRKWZhaCOlmS4M84YW9E2wvAuav7RgQn492Vhv3Ksg_IUfrQFkZs5ZrPfHSbGQZ5gWDLYnI-hQ), which makes the argument that (irrelevant) emotions during learning can inflate perceived learning. More generally, research on emotions and memory (e.g., flashbulb memories) could inform the hypotheses for the role of emotional intensity in the proposed study.

[Reviewer 3] The proposal is imprecise with the terminology and lacks precise construct definitions. Is emotional intensity the same as emotional involvement and self-involvement? Whereas the first may relate to valence and arousal as the two fundamental dimensions of emotional experience, the third one is more connected to interest, previous experience, or other more cognitive factors. Perhaps also because of the lack of theoretical context, it does not become clear which construct the proposed manipulation should actually target.

[Reviewer 3] The authors write: “Following Park’s intuition (2001) we believe that the key characteristic that might inflate perceived knowledge is the perceived involvement of the individual, regardless of the topic being assessed.” (p.4) But why do the authors “believe” that? I could not find any theoretical explanation in the proposal. The authors further propose two hypotheses H2 and H4 on emotional involvement without a theoretical explanation of why emotional involvement should have the proposed effects. I am sorry if I missed something here; I found the design and hypotheses lacked a clear theoretical fundament and sufficient explanation.

Authors’ response:

We initially thought "emotional intensity" would be a straightforward label for the classification of the topics, but clearly, we were mistaken. We thank the reviewers for pointing that out. The key factor characterizing the chosen topics is self-involvement, derived from averaging responses to "how emotionally

involved do you feel?" and "How willing are you to discuss the topic" from the pre-test (Appendix A). Since these two answers showed a strong correlation (Cronbach's Alpha = 0.87, 0.859 - 0.88 [CI .99]), we combined them into a single self-involvement construct, aligning with Park's protocol. We now call this index simply "self-involvement" to avoid any mistake in interpretation. We have gone through the manuscript and changed the terminology accordingly, and also tried to make the justification for the hypothesis more clear.

[Reviewer 2] The question about emotional involvement can be said to be a manipulation check for the emotional intensity variable. Here, one would obviously predict higher involvement for high intensity than for low intensity topics. Similarly, the baseline knowledge scores would presumably also differ between low and high knowledge topics. It would be good to specify these points in the manuscript.

[Reviewer 3] The authors state "If so, it may be that our selected topics failed to emotionally involve to the right extent, or, that emotional intensity does not have an effect per se. " (p.8). However, this is what a manipulation check could reveal, which the authors actually plan to assess by measuring the emotional involvement.

Authors' response:

We thank the reviewers again for this very helpful consideration. We included a paragraph in the manuscript (page 11) that presents the above manipulation checks as a new hypothesis (H0) and as an exploratory analysis: specifically, we will test whether self-involvement by participants correlates with self-involvement by participants in the pre-test (H0), and similarly, perceived knowledge by participants correlates with perceived knowledge in the pre-test.

Knowledge illusion measures

[Reviewer 2] There are some inconsistencies for the illusion of knowledge measure. The illusion of knowledge is stated to be calculated as "the difference between perceived knowledge at T2 and actual knowledge, that is the proportion of correct answers: ki = pkT2 – score of factual knowledge". Perceived knowledge is measured using scale from 1 (nothing) to 100 (everything). Factual knowledge is measured as the proportion of correct answers, and so goes from 0 to 1.
To make the illusion of knowledge measure more meaningful, I think some changes need to be made. First, the perceived knowledge scale should go from 0 to 100, so that the bounds are

similar between perceived and factual knowledge. As of now, a 0 score is possible for factual but not for perceived knowledge. Second, and more importantly, the two measures should both go from 0 to 100 or from 0 to 1. Otherwise, it will be harder to interpret the illusion of knowledge measure (e.g., someone who scored 50 on perceived knowledge and had 5 correct questions would receive an illusion of knowledge score of 49.5). I think converting the factual knowledge measure to a 0 to 100 scale makes most sense.

[Reviewer 3] the authors want to operate with difference scores as dependent variable. I strongly advise against this because this eliminates all main effects of emotional intensity a priori. For example, 5-3 is treated equally to 7-5. Yet, the different baseline levels between conditions may be theoretically relevant. (Also, they might reveal floor/ceiling effects for some conditions that are otherwise not detectable.)

Authors' response:

Thank you for your constructive suggestions. We agree with the points you raised and have decided to normalize the scales of our measurements of perceived and factual knowledge to a range from 0 to 100. We also have now included an explanation in the manuscript about the standardization. Our approach will involve standardizing all data and transforming them into a uniform 0 to 1 range. For example, participants who scored 50 on perceived knowledge will receive a score of 0.5. If they had 5 correct answers, their actual knowledge score would be 0.5, resulting in an illusion of knowledge score of 0, as they accurately assessed their level of knowledge.

Control questions

[Reviewer 2] it could be good to include a manipulation check for exposure, for example by asking (after completion of other measures) which of the topics the participant (remembers) reading about in the experiment. There may be better ways to include some positive control for news exposure, but the authors should at least consider whether and how they could do this.

[Reviewer 2] The authors also note (p. 7) that "Some extra control questions will be administered to check whether subjects had paid attention to the experimental stimuli and environment". It would be good to specify what these control questions were, and whether they were administered at T1, at T2, or both.

[Reviewer 3] The authors state "Some extra control questions will be administrated to check whether subjects had paid attention to the experimental stimuli and environment." Please be more transparent regarding these attention checks to ensure reproducibility. The proposal

describes the specific control variables they assess but does not outline a specific analytical strategy for these variables.

Authors' response:

Our description of attention checks was indeed lacking in the manuscript, thank you for bringing up this point. We decided to include both an attention and a manipulation check. At T1, during the psychometric assessment, we will include items to test whether the participant is reading the questions rather than clicking randomly (attention check). An example of these items is:

"Please answer "Totally disagree" to this question"

At T2, to ensure that participants have been attentive while scrolling through the newsfeed, we will ask them to recall the topics they remember. In particular, we will ask them if they remember to have seen news about two topics, one actually belonging to their experimental group news feed, and one randomly taken from the other treatments.

Participants who fail these checks will not be excluded from the sample: we will test our hypothesis with and without them (as a robustness exploratory analysis) to assess the impact of their answers on the data.

Language of the materials

[Reviewer 2] in the appendices, there is a mix of English and Italian when it comes to measures and topics. For better replicability, I think all materials should be available in English.

[Reviewer 3] I appreciate that the original articles and the items of the knowledge tests are included in the proposal. However, as a non-Italian, I cannot provide any feedback here.

Authors' response:

The English translation has been provided next to the Italian wording. It is important to point out, however, that the topics and, therefore, the knowledge assessments have been calibrated on an Italian sample. Thank you for highlighting it.

Minor corrections

[Reviewer 1] Appendix C: the links are not clickable.

We fixed the issue.

[Reviewer 1] Appendix d: It would have been great to have the Likert scales (1 to 5 / 1 to 7 …). They are not explained in the procedure also so they might at least have been said here.

We included the description of the likert scales used in Appendix D.

[Reviewer 2] for Hypothesis 3 and 4, the term “ki” is used in the equations, as an abbreviation of illusion of knowledge. However, this term is only defined two pages later, in the description of the measures. Please introduce this term together with the equations to improve comprehension.

We specified the meaning of the abbreviation.

[Reviewer 3] p. 2: “As far as we are aware, only two empirical studies” → Would it be more accurate to speak of “experimental studies” here, given the correlational evidence mentioned by the authors?

We changed the wording according to the kind suggestion.

[Reviewer 3] p.4 “Both experiments were implemented as between-subjects designs where participants were first exposed to a newsfeed or a news article and then asked about their perceived and factual knowledge. […] The results indicated that participants who scrolled through many article previews had a significantly higher perceived knowledge that did not match their actual knowledge. “→ I did not fully comprehend the specific design and the corresponding comparison to arrive at this statement. It might help mentioning the design of these two articles here (e.g., experimental vs. control condition) and to be more precise what the comparative statement (“higher perceived knowledge”) refers to as a comparison standard.

We specified further the experimental protocol of both experiments, as the control conditions were different.

[Reviewer 3] p. 4 “Consequently, without a pre-test, the estimation of perceived knowledge obtained after an actual knowledge test may be biased by this intervention.” I did not fully comprehend what type of bias the authors meant here (e.g., an underestimation of the effect).

We included a specification of the expected effect.

[Reviewer 3] p. 4 What is “perceived involvement of the individual”? – who is the perceiver here?

We rephrased it to improve the clarity of the sentence.

[Reviewer 3] p. 4 The authors discuss limitations of Anspach et al. (2019) as one motivation for their own research, but no limitations of Schäfer (2020). I was just a little confused because I expected it after the limitations of Anspach.

> It is not possible to infer similar limitations from the protocol described by Schäfer, so we decided to report one of the limitations listed by Schäfer herself that we aim to overcome with our study, which is the use of many experimental topics.

[Reviewer 3] p. 8 → “Once the experiment is ready to run, Prolific will send an invitation email to all potential participants” → I did not know that it was possible to invite participants per mail on Prolific. Could the authors share how they did that (i.e., whether this is some custom allowlist or some other function that allows this)?

> Prolific sends an invitation email to potential participants who meet the requirements set by the researchers, we will not invite any participants as we do not have access to any of their data.