THE ROYAL SOCIETY
PUBLISHING

PROCEEDINGS B

Dogs follow human misleading suggestions more often when the informant has a false belief

Lucrezia Lonardo, Christoph Völter, Claus Lamm and Ludwig Huber

Article citation details
Proc. R. Soc. B **288**: 20210906.
http://dx.doi.org/10.1098/rspb.2021.0906

Review timeline

Original submission:	17 April 2021
Revised submission:	22 June 2021
Final acceptance:	25 June 2021

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSPB-2021-0906.R0 (Original submission)

Review form: Reviewer 1

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
Through a novel cue-following object choice task, the authors provide the first evidence of dogs' sensitivity to others' beliefs. The study is of pronounced theoretical significance, and it is pre-registered, thoughtfully designed, unusually well-powered, clear and well-written. It includes a series of three studies within which the key results are clarified and replicated and an important alternative explanation is controlled for. Importantly, the authors do not attempt to over-interpret their results, clearly pointing (in the paper's second to last sentence) to several mechanistic hypotheses that can be tested in future research. The ability to track others' perspectives – and particularly others' beliefs – is at the heart of human sociality, from culture to cooperation to language. And consequently, for over forty years, researchers have been interested in whether this capacity is unique to humans. The authors provide the first evidence of potential false belief representation outside of primates. This paper is sure to garner substantial interest and inspire further research into the mechanisms and cognitive representations that support dogs' success in this task. It will also surely lead to investigations of the selective and environmental forces that produced convergent capacities in primates and dogs (and maybe other species, like corvids), and that resulted in divergent performance across breed classes. It is exceedingly rare to review a paper that is so strong and so obviously deserving of publication in nearly its current format. Well done!

In my view, the authors have interpreted their results fairly in the discussion, including attempting to understand the unpredicted direction of their primary effects and pointing to alternative accounts for future research. The discussion could be further bolstered, however, by at least briefly addressing open questions about whether dogs' capacities are likely to be shared with wolves (or not) and whether they are likely to result from convergent evolution with primates (and maybe corvids) or shared ancestry. Relatedly, it could be valuable to highlight the need for future work that precisely characterizes the mechanisms across species, to determine the extent to which identical (as opposed to only superficially similar) mechanisms have evolved (perhaps convergently).

Minor points:

Line 39: this is true but the task was actually proposed by Dennett in his 1978 commentary "Beliefs about beliefs" in response to Premack and Woodruff's target article.

Line 64: note that Heyes' submentalizing hypothesis has been directly tested in apes:
Kano et al 2017 Submentalizing cannot explain belief-based action anticipation in apes. Trends

Cogn Sci
Krupenye et al 2017 A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. Communicative & Integrative Biology

And in human infants:
Surian and Franchin 2020 On the domain specificity of the mechanisms underpinning spontaneous anticipatory looks in false-belief tasks. Dev Sci

Lines 176-215: when reading the procedure (i.e., before arriving to the results section), it was not clear which conditions were performed within each experiment. It would be helpful to make this explicit already in the procedure section for comprehension of the procedure and scoring and analysis sections.

Scoring and analysis: If space permits, I think it is worth moving the inter-rater reliability information to the main text

Procedure/scoring: I did not see information in the main text about (1) maximum trial length before it would be scored as no choice, or (2) whether trials were repeated in the event of no choice

Lines 225-227: on first read, I misread the passage as indicting that all mentioned variables were included as random intercepts (which of course would not be possible). Separating fixed effects and random effects into separate sentences could ensure no confusion here.

Scoring and analysis: The analysis strategy is sound but usually, when pursuing null hypothesis significance testing, a null model is created with only random effects and control variables (i.e., with test predictors removed) and null and full models are compared with a likelihood ratio test before the full model is interpreted. In the authors' case, their full models only include a single test predictor (experimental condition), meaning that the full-null comparison is exactly the same as the drop1 comparison used to generate the p-value for the test predictor. Maybe it is worth explicitly stating this for readers who will be wondering where the null model (and full-null model comparison) is?

Analysis: If space permits, the authors should describe in the main text how significance of individual predictors was determined (drop1 function)

Analysis: The authors should also describe how pairwise comparisons were performed in cases when they did not derive directly from the models (e.g., in the model with all three conditions, were pairwise comparisons produced from re-levelling the variables or another function or were they based on other techniques distinct from the models themselves?

Line 237: From the analysis section, I had the impression that Experiments 1 and 2 were analyzed in a single GLMM that included all 3 conditions. However, it becomes less clear in the results section whether all three conditions were analyzed together or whether the data from Experiment 1 were initially analyzed on their own. The supplement is clearer in this respect but this issue could be clarified in the main text.

Review form: Reviewer 2

Recommendation
Major revision is needed (please make suggestions in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
I really enjoyed reading this manuscript on human belief attribution by dogs. This is timely work and important results on belief attribution in a non-primate species. The results will very likely spark the general discussion on ToM-like skills in non-human species.

The authors tested dogs in two groups: one group observed a human pointing to a previously baited location, holding a true belief, while the other group observed a human holding a false belief. Using a preregistered study design for Experiment 1, they found that dogs in both groups differed in their behaviour – but in the opposite direction than the authors expected. After visual inspection of breed-specific variances in the choice data, the authors decided to run a follow-up study comparing border collies and terriers, where they could statistically support a difference in the group of terriers that is in line with their preregistered hypothesis. I am sincerely impressed by the comprehensive and detailed provision of behavioural data that made it very easy to follow the pre-defined analysis steps chosen by the authors (including open script, code, exploratory analysis and additional relevant statistics in the ESM).
However, I have a few reservations regarding the interpretation of the data. One of my main concerns is about the rationale of Experiment 3 and the subsequent discussion of the different findings in Exp 1-2 and 3. I fully agree that breed-specific differences warrant further attention (and the authors themselves speculate in the discussion about a potential difference of more cooperative breeds vs more independent breeds). I am thus surprised that the authors only chose one specific FCI group to follow up on their initial results. In my opinion, this selection was

premature and hampers, rather than advances, the discussion on the topic.
The authors state that terriers were the only ones that have chosen container A than B more often, while other groups showed no difference or the opposite pattern (286-287) – for me, it is not clear what criterion was used to make this inference (as e.g. FCI group 1 also shows a tendency to choose container A more often).
In addition, group 3 (terriers) only consisted of 10 subjects, and, if I am interpreting the wideness of the bars correctly, only 3 or 4 subjects were tested in the TB group – which makes it almost impossible to detect whether this pattern can be statistically supported. In other words: as there were also other FCI groups with samples between 10-15 subjects, having 2-3 animals chosen differently in the TB group could have already made a difference in the visual inspection of the graph and the subsequent selection for the follow-up study.
Given their data from Experiment 1 and their discussion on potential differences between more cooperative vs more independent breeds, why haven't the authors at least run an exploratory analysis from their initial dataset to potentially support their alternative explanation?
I was additionally wondering why the authors in Exp3 have not chosen a range of breeds from FCI 1 and opted for one single breed instead – in particular, as the authors did not focus on one breed of terriers.
I wanted to emphasize that the authors might consider also to discuss that a sample of dogs that is prone to the perseveration bias (A not B error) has likely been excluded from the tested population (following Fam phase 1). As performance on this task could potentially be linked to difference in attention (to the task) and/or gathering information based on local enhancement, it would be interesting to know how many dogs from each FCI group had to be excluded from the test based on what criterion.
On a last general remark, I think that the introduction and discussion are in parts quite lengthy, while other parts need some additional crucial information to provide the reader with a better walk-through of the study rationale (see detailed comments).
Detailed comments:
General
Please provide an ethical approval number
Intro
Why would we expect differences in a cooperative and a competitive task? How does a competitive task look like? Given that the authors have chosen a cooperative task, this appears like crucial background to understand the study rationale.
38 a little bit more background on the importance of the seminal article would be advantageous for readers unfamiliar with the field
40 a bracket is missing at the end of the sentence
48 please add the Latin name for chimps
58-64 I think it would add to the comprehension of these criticisms if the authors would suggest a potential solution to them
66 here the authors mention that dogs are a particularly interesting case because of their shared social environment – as this would likely also be the case for the common housefly, I think the authors should elaborate why dogs are good human behaviour readers (and might constitute a different case than, e.g., other domestic animal species)
74-109 I think a lot of the information covered here should actually be placed in the method section (e.g. 79-82; 89-92)
114 the authors should elaborate here what they mean with retroactive interference
120-137 I think that some details here could be cut down (and moved to the discussion)
137 Reference(s) missing
Methods
Familiarisation phase 3: "Only dogs that made two correct choices in two consecutive trials (one with displacement, one without displacement) within four trials were subsequently tested in the final test phase." Why has this rather weak criterion been chosen?
What posthoc test was used for the analysis of Exp1-2?
151-152 please provide more details – why match with TB, but not FB group?
Discussion
282-292 Summarising the rationale in the first paragraph of the discussion is a great way to

remind the reader about the key questions that the authors aimed to answer – I would additionally favour that the authors would also include a very brief summary of the results (and the inferences that they can draw from them) in this paragraph
293-310 I think these two paragraphs should be merged. I also think that prior to discussing the control condition from Exp2, the results of Exp1 should be discussed in a bit more detail
311-12 the authors should elaborate here why this result was surprising, e.g. by stating the directionality of their hypothesis prior to this statement
331-333: even if dogs perceive the misleading pointing in the FB group as a mistake in goodwill (or any mistake for whatever reason), they should still see it as a mistake and choose the baited container, right? Or are the authors aiming to make the case that the human in the FB group is still be seen as a collaborator (rather than the one in the TB group -> deceiver) and that this difference in perceived trustworthiness gives dogs a higher inclination to be misled? What would be the mechanisms that would cause such an inclination?
348 I think it would be very informative to also have the number of subjects stated that were assigned to each treatment group from each FCI group (e.g., in the ESM). Given the difference in numbers between FCI groups (and FCI groups x treatment groups) it might be better to change phrases such as "more terriers" to "relatively more terriers" etc.
362-365 Why would this specifically be the case for terriers (see also the main comment on why terriers have been chosen for exp3)?
ESM
"For Experiment 3, we conducted another power simulation to determine the sample size. This revealed a power of 76% with 40 dogs and an expected performance of 0.3 in the FB group and 0.7 in the TB group (performance predictions based on the terriers performance in Experiment 1).". As the performance of terriers was about 0.2 in the FB group and 1.0 in the TB group, I assume that the authors choose more conservative numbers based on the low sample size in Exp1?
Did the authors observe breed differences in anticipatory looking? I am just wondering what were the reasons the authors opted to not include it (N is likely to low?), as all FCI group data has been plotted for choice and latency, but not anticipatory behaviour
Plot width differs in e.g. Figure ESM S2 – I cannot find information in the legend, but might assume that the bar widths represents the sample size for each treatment (similar to the other figures)?

Review form: Reviewer 3 (Peter Pongracz)

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
No

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
Review on the manuscript
Dogs follow human misleading suggestions more often when the informant has a false belief
Written by Lonardo et al.

Overall opinion
This is an interesting and well-written study, where authors experimentally tested whether dogs act differently in a visible food displacement two-choice task when they are supported by either a knowledgeable or non-knowledgeable human informant (the 'communicator'). As the informant always suggested to the dogs to choose the 'wrong' container (where the food was moved from to the other container), the information can be regarded as false. Thus, dogs' eventual choice patterns can be regarded as whether they were dependent on the 'false belief' on the non-knowledgeable informant.
The experiments are carefully designed and analyzed, they were run on an impressive size sample. The results are interesting and their interpretation is modest. Authors carefully took care of the important alternative hypotheses.
My critical comments are mostly minor (listed below). I found two issues that can represent some level of concern. One is the inclusion of rather young (5-12 months of age) subjects to the sample. The other is a somewhat tricky question: can we be sure that dogs would chose the 'good' container without any intervention of the 'communicator'? Authors should provide a good reasoning for this, unless they do not want to include one more control group (which I doubt, understandably).

Detailed comments
I would recommend to use an important piece of literature that is one of the rare attempts in the past to test dogs' performance in a task where a human 'helper' had correct or incorrect knowledge about the task (hence, this paper can be regarded as testing for 'false belief' attribution in dogs):
Virányi, Z., Topál, J., Miklósi, Á., & Csányi, V. (2006). A nonverbal test of knowledge attribution: a comparative study on dogs and children. Animal Cognition, 9(1), 13-26.
The manuscript has a very odd, somewhat confusing structure. After the true Introduction comes a long section (lines 74-137), which is formally still within the Intro, but in reality it is a mixture of methods and hypotheses. I do believe that the research goal, questions, predictions and hypotheses should be placed to the end of the Introduction, but in this case it was done in a too large extent. Especially the detailed methodology seems like a repetition, because later in the Methods this is done again.

ould you provide a justification for the inclusion of both juvenile (5-12 months) and adult dogs to the test population? Although there are results that show juvenile (or even younger) dogs performing comparably well in tasks that involve inter-specific social cognition, there can be factors that affect differently the performance of juvenile and adult dogs (level of training, attention span, inhibitory threshold etc.).
How long was the pause between baiting bucket A and transferring the food to bucket B in case of the three initial familiarization trials? Was it comparably long to the pause that they used in the testing trials, or was it shorter?
What sort of food reward was used for baiting? Were the pieces large enough for the dogs to see their transfer to and from the buckets?
Did dogs always choose the bucket that was suggested by the communicator in the warm-up trials?
Line 269 – please add the names of the particular FCI dog breed groups here to the results. Referring to their FCI-numbers only, hides the interesting nature of the results, which shows remarkable pattern according to the breed groups.
Line 328 – Authors here state that "Along this line of argument, dogs in both groups remembered the final location of food (bucket B)." Actually, we cannot know this, because there was no control group that would test dogs' choices WITHOUT the interference of the 'communicator' experimenter. In such a control group, everything would happen as in the False Belief group, but the 'communicator' would not suggest any of the buckets for the dogs before those are let to make their choice.

Decision letter (RSPB-2021-0906.R0)

@@date to be populated upon sending@@

Dear Miss Lonardo:

Your manuscript has now been peer reviewed and the reviews have been assessed by an Associate Editor. The reviewers' comments (not including confidential comments to the Editor) and the comments from the Associate Editor are included at the end of this email for your reference. As you will see, the reviewers and the Editors have raised some concerns with your manuscript and we would like to invite you to revise your manuscript to address them.

We do not allow multiple rounds of revision so we urge you to make every effort to fully address all of the comments at this stage. If deemed necessary by the Associate Editor, your manuscript will be sent back to one or more of the original reviewers for assessment. If the original reviewers are not available we may invite new reviewers. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

To submit your revision please log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions", click on "Create a Revision". Your manuscript number has been appended to denote a revision.

When submitting your revision please upload a file under "Response to Referees" - in the "File Upload" section. This should document, point by point, how you have responded to the reviewers' and Editors' comments, and the adjustments you have made to the manuscript. We require a copy of the manuscript with revisions made since the previous version marked as 'tracked changes' to be included in the 'response to referees' document.

Your main manuscript should be submitted as a text file (doc, txt, rtf or tex), not a PDF. Your figures should be submitted as separate files and not included within the main manuscript file.

When revising your manuscript you should also ensure that it adheres to our editorial policies (https://royalsociety.org/journals/ethics-policies/). You should pay particular attention to the following:

Research ethics:
If your study contains research on humans please ensure that you detail in the methods section whether you obtained ethical approval from your local research ethics committee and gained informed consent to participate from each of the participants.

Use of animals and field studies:
If your study uses animals please include details in the methods section of any approval and licences given to carry out the study and include full details of how animal welfare standards were ensured. Field studies should be conducted in accordance with local legislation; please include details of the appropriate permission and licences that you obtained to carry out the field work.

Data accessibility and data citation:
It is a condition of publication that you make available the data and research materials supporting the results in the article. Please see our Data Sharing Policies (https://royalsociety.org/journals/authors/author-guidelines/#data). Datasets should be deposited in an appropriate publicly available repository and details of the associated accession number, link or DOI to the datasets must be included in the Data Accessibility section of the article (https://royalsociety.org/journals/ethics-policies/data-sharing-mining/). Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available).

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should also be fully cited and listed in the references.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link http://datadryad.org/submit?journalID=RSPB&manu=(Document not available), which will take you to your unique entry in the Dryad repository.

If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link.

For more information please see our open data policy http://royalsocietypublishing.org/data-sharing.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI. Please try to submit all supplementary material as a single file.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

Please submit a copy of your revised paper within three weeks. If we do not hear from you within this time your manuscript will be rejected. If you are unable to meet this deadline please let us know as soon as possible, as we may be able to grant a short extension.

Thank you for submitting your manuscript to Proceedings B; we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes,
Dr Robert Barton
mailto: proceedingsb@royalsociety.org

Associate Editor
Comments to Author:
Three reviewers have provided feedback on this article. All three highlight the importance of this article in expanding the focus of false belief understanding research beyond the study of primates. Furthermore, the reviewers acknowledge the value of pre-registration and the sound methodological design employed in this study. While all three reviewers are generally very favorable towards this study, all three provide detailed and thoughtful feedback as to how the authors can further clarify and enhance their reporting and I agree with the suggestions made. Furthermore, reviewer 2 requests a more critical appraisal of the results of experiment 3 and proposes alternative ways to consider the data more generally. I think such thoroughness would benefit the robustness of the conclusions drawn.

Reviewer(s)' Comments to Author:
Referee: 1
Comments to the Author(s)
Through a novel cue-following object choice task, the authors provide the first evidence of dogs' sensitivity to others' beliefs. The study is of pronounced theoretical significance, and it is pre-registered, thoughtfully designed, unusually well-powered, clear and well-written. It includes a series of three studies within which the key results are clarified and replicated and an important alternative explanation is controlled for. Importantly, the authors do not attempt to over-interpret their results, clearly pointing (in the paper's second to last sentence) to several mechanistic hypotheses that can be tested in future research. The ability to track others' perspectives – and particularly others' beliefs – is at the heart of human sociality, from culture to cooperation to language. And consequently, for over forty years, researchers have been interested in whether this capacity is unique to humans. The authors provide the first evidence of potential false belief representation outside of primates. This paper is sure to garner substantial interest and inspire further research into the mechanisms and cognitive representations that support dogs' success in this task. It will also surely lead to investigations of the selective and environmental forces that produced convergent capacities in primates and dogs (and maybe other species, like corvids), and that resulted in divergent performance across breed classes. It is exceedingly rare to review a paper that is so strong and so obviously deserving of publication in nearly its current format. Well done!

In my view, the authors have interpreted their results fairly in the discussion, including attempting to understand the unpredicted direction of their primary effects and pointing to alternative accounts for future research. The discussion could be further bolstered, however, by at least briefly addressing open questions about whether dogs' capacities are likely to be shared with wolves (or not) and whether they are likely to result from convergent evolution with primates (and maybe corvids) or shared ancestry. Relatedly, it could be valuable to highlight the need for future work that precisely characterizes the mechanisms across species, to determine the extent to which identical (as opposed to only superficially similar) mechanisms have evolved (perhaps convergently).

Minor points:

Line 39: this is true but the task was actually proposed by Dennett in his 1978 commentary "Beliefs about beliefs" in response to Premack and Woodruff's target article.

Line 64: note that Heyes' submentalizing hypothesis has been directly tested in apes:
Kano et al 2017 Submentalizing cannot explain belief-based action anticipation in apes. Trends Cogn Sci
Krupenye et al 2017 A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. Communicative & Integrative Biology

And in human infants:
Surian and Franchin 2020 On the domain specificity of the mechanisms underpinning spontaneous anticipatory looks in false-belief tasks. Dev Sci

Lines 176-215: when reading the procedure (i.e., before arriving to the results section), it was not clear which conditions were performed within each experiment. It would be helpful to make this explicit already in the procedure section for comprehension of the procedure and scoring and analysis sections.

Scoring and analysis: If space permits, I think it is worth moving the inter-rater reliability information to the main text

Procedure/scoring: I did not see information in the main text about (1) maximum trial length before it would be scored as no choice, or (2) whether trials were repeated in the event of no choice

Lines 225-227: on first read, I misread the passage as indicting that all mentioned variables were included as random intercepts (which of course would not be possible). Separating fixed effects and random effects into separate sentences could ensure no confusion here.

Scoring and analysis: The analysis strategy is sound but usually, when pursuing null hypothesis significance testing, a null model is created with only random effects and control variables (i.e., with test predictors removed) and null and full models are compared with a likelihood ratio test before the full model is interpreted. In the authors' case, their full models only include a single test predictor (experimental condition), meaning that the full-null comparison is exactly the same as the drop1 comparison used to generate the p-value for the test predictor. Maybe it is worth explicitly stating this for readers who will be wondering where the null model (and full-null model comparison) is?

Analysis: If space permits, the authors should describe in the main text how significance of individual predictors was determined (drop1 function)

Analysis: The authors should also describe how pairwise comparisons were performed in cases when they did not derive directly from the models (e.g., in the model with all three conditions, were pairwise comparisons produced from re-levelling the variables or another function or were they based on other techniques distinct from the models themselves?

Line 237: From the analysis section, I had the impression that Experiments 1 and 2 were analyzed in a single GLMM that included all 3 conditions. However, it becomes less clear in the results section whether all three conditions were analyzed together or whether the data from Experiment 1 were initially analyzed on their own. The supplement is clearer in this respect but this issue could be clarified in the main text.

Referee: 2
Comments to the Author(s)
I really enjoyed reading this manuscript on human belief attribution by dogs. This is timely work and important results on belief attribution in a non-primate species. The results will very likely spark the general discussion on ToM-like skills in non-human species.

The authors tested dogs in two groups: one group observed a human pointing to a previously baited location, holding a true belief, while the other group observed a human holding a false belief. Using a preregistered study design for Experiment 1, they found that dogs in both groups differed in their behaviour – but in the opposite direction than the authors expected. After visual inspection of breed-specific variances in the choice data, the authors decided to run a follow-up study comparing border collies and terriers, where they could statistically support a difference in the group of terriers that is in line with their preregistered hypothesis. I am sincerely impressed by the comprehensive and detailed provision of behavioural data that made it very easy to follow the pre-defined analysis steps chosen by the authors (including open script, code, exploratory analysis and additional relevant statistics in the ESM).

However, I have a few reservations regarding the interpretation of the data. One of my main concerns is about the rationale of Experiment 3 and the subsequent discussion of the different findings in Exp 1-2 and 3. I fully agree that breed-specific differences warrant further attention (and the authors themselves speculate in the discussion about a potential difference of more cooperative breeds vs more independent breeds). I am thus surprised that the authors only chose one specific FCI group to follow up on their initial results. In my opinion, this selection was premature and hampers, rather than advances, the discussion on the topic.

The authors state that terriers were the only ones that have chosen container A than B more often, while other groups showed no difference or the opposite pattern (286-287) – for me, it is not clear what criterion was used to make this inference (as e.g. FCI group 1 also shows a tendency to choose container A more often).

In addition, group 3 (terriers) only consisted of 10 subjects, and, if I am interpreting the wideness of the bars correctly, only 3 or 4 subjects were tested in the TB group – which makes it almost impossible to detect whether this pattern can be statistically supported. In other words: as there were also other FCI groups with samples between 10-15 subjects, having 2-3 animals chosen differently in the TB group could have already made a difference in the visual inspection of the graph and the subsequent selection for the follow-up study.

Given their data from Experiment 1 and their discussion on potential differences between more cooperative vs more independent breeds, why haven't the authors at least run an exploratory analysis from their initial dataset to potentially support their alternative explanation?

I was additionally wondering why the authors in Exp3 have not chosen a range of breeds from FCI 1 and opted for one single breed instead – in particular, as the authors did not focus on one breed of terriers.

I wanted to emphasize that the authors might consider also to discuss that a sample of dogs that is prone to the perseveration bias (A not B error) has likely been excluded from the tested population (following Fam phase 1). As performance on this task could potentially be linked to difference in attention (to the task) and/or gathering information based on local enhancement, it would be interesting to know how many dogs from each FCI group had to be excluded from the test based on what criterion.

On a last general remark, I think that the introduction and discussion are in parts quite lengthy, while other parts need some additional crucial information to provide the reader with a better walk-through of the study rationale (see detailed comments).

Detailed comments:
General
Please provide an ethical approval number

Intro

Why would we expect differences in a cooperative and a competitive task? How does a competitive task look like? Given that the authors have chosen a cooperative task, this appears like crucial background to understand the study rationale.

38 a little bit more background on the importance of the seminal article would be advantageous for readers unfamiliar with the field

40 a bracket is missing at the end of the sentence

48 please add the Latin name for chimps

58-64 I think it would add to the comprehension of these criticisms if the authors would suggest a potential solution to them

66 here the authors mention that dogs are a particularly interesting case because of their shared social environment – as this would likely also be the case for the common housefly, I think the authors should elaborate why dogs are good human behaviour readers (and might constitute a different case than, e.g., other domestic animal species)

74-109 I think a lot of the information covered here should actually be placed in the method section (e.g. 79-82; 89-92)

114 the authors should elaborate here what they mean with retroactive interference

120-137 I think that some details here could be cut down (and moved to the discussion)

137 Reference(s) missing

Methods

Familiarisation phase 3: "Only dogs that made two correct choices in two consecutive trials (one with displacement, one without displacement) within four trials were subsequently tested in the final test phase." Why has this rather weak criterion been chosen?

What posthoc test was used for the analysis of Exp1-2?

151-152 please provide more details – why match with TB, but not FB group?

Discussion

282-292 Summarising the rationale in the first paragraph of the discussion is a great way to remind the reader about the key questions that the authors aimed to answer – I would additionally favour that the authors would also include a very brief summary of the results (and the inferences that they can draw from them) in this paragraph

293-310 I think these two paragraphs should be merged. I also think that prior to discussing the control condition from Exp2, the results of Exp1 should be discussed in a bit more detail

311-12 the authors should elaborate here why this result was surprising, e.g. by stating the directionality of their hypothesis prior to this statement

331-333: even if dogs perceive the misleading pointing in the FB group as a mistake in goodwill (or any mistake for whatever reason), they should still see it as a mistake and choose the baited container, right? Or are the authors aiming to make the case that the human in the FB group is still be seen as a collaborator (rather than the one in the TB group -> deceiver) and that this difference in perceived trustworthiness gives dogs a higher inclination to be misled? What would be the mechanisms that would cause such an inclination?

348 I think it would be very informative to also have the number of subjects stated that were assigned to each treatment group from each FCI group (e.g., in the ESM). Given the difference in numbers between FCI groups (and FCI groups x treatment groups) it might be better to change phrases such as "more terriers" to "relatively more terriers" etc.

362-365 Why would this specifically be the case for terriers (see also the main comment on why terriers have been chosen for exp3)?

ESM

"For Experiment 3, we conducted another power simulation to determine the sample size. This revealed a power of 76% with 40 dogs and an expected performance of 0.3 in the FB group and 0.7 in the TB group (performance predictions based on the terriers performance in Experiment 1).". As the performance of terriers was about 0.2 in the FB group and 1.0 in the TB group, I assume that the authors choose more conservative numbers based on the low sample size in Exp1?

Did the authors observe breed differences in anticipatory looking? I am just wondering what were the reasons the authors opted to not include it (N is likely to low?), as all FCI group data has been plotted for choice and latency, but not anticipatory behaviour

Plot width differs in e.g. Figure ESM S2 – I cannot find information in the legend, but might assume that the bar widths represents the sample size for each treatment (similar to the other figures)?

Referee: 3
Comments to the Author(s)
Review on the manuscript
Dogs follow human misleading suggestions more often when the informant has a false belief
Written by Lonardo et al.

Overall opinion
This is an interesting and well-written study, where authors experimentally tested whether dogs act differently in a visible food displacement two-choice task when they are supported by either a knowledgeable or non-knowledgeable human informant (the 'communicator'). As the informant always suggested to the dogs to choose the 'wrong' container (where the food was moved from to the other container), the information can be regarded as false. Thus, dogs' eventual choice patterns can be regarded as whether they were dependent on the 'false belief' on the non-knowledgeable informant.
The experiments are carefully designed and analyzed, they were run on an impressive size sample. The results are interesting and their interpretation is modest. Authors carefully took care of the important alternative hypotheses.
My critical comments are mostly minor (listed below). I found two issues that can represent some level of concern. One is the inclusion of rather young (5-12 months of age) subjects to the sample. The other is a somewhat tricky question: can we be sure that dogs would chose the 'good' container without any intervention of the 'communicator'? Authors should provide a good reasoning for this, unless they do not want to include one more control group (which I doubt, understandably).

Detailed comments
I would recommend to use an important piece of literature that is one of the rare attempts in the past to test dogs' performance in a task where a human 'helper' had correct or incorrect

knowledge about the task (hence, this paper can be regarded as testing for 'false belief' attribution in dogs):
Virányi, Z., Topál, J., Miklósi, Á., & Csányi, V. (2006). A nonverbal test of knowledge attribution: a comparative study on dogs and children. Animal Cognition, 9(1), 13-26.
The manuscript has a very odd, somewhat confusing structure. After the true Introduction comes a long section (lines 74-137), which is formally still within the Intro, but in reality it is a mixture of methods and hypotheses. I do believe that the research goal, questions, predictions and hypotheses should be placed to the end of the Introduction, but in this case it was done in a too large extent. Especially the detailed methodology seems like a repetition, because later in the Methods this is done again.
ould you provide a justification for the inclusion of both juvenile (5-12 months) and adult dogs to the test population? Although there are results that show juvenile (or even younger) dogs performing comparably well in tasks that involve inter-specific social cognition, there can be factors that affect differently the performance of juvenile and adult dogs (level of training, attention span, inhibitory threshold etc.).
How long was the pause between baiting bucket A and transferring the food to bucket B in case of the three initial familiarization trials? Was it comparably long to the pause that they used in the testing trials, or was it shorter?
What sort of food reward was used for baiting? Were the pieces large enough for the dogs to see their transfer to and from the buckets?
Did dogs always choose the bucket that was suggested by the communicator in the warm-up trials?
Line 269 – please add the names of the particular FCI dog breed groups here to the results. Referring to their FCI-numbers only, hides the interesting nature of the results, which shows remarkable pattern according to the breed groups.
Line 328 – Authors here state that "Along this line of argument, dogs in both groups remembered the final location of food (bucket B)." Actually, we cannot know this, because there was no control group that would test dogs' choices WITHOUT the interference of the 'communicator' experimenter. In such a control group, everything would happen as in the False Belief group, but the 'communicator' would not suggest any of the buckets for the dogs before those are let to make their choice.

Author's Response to Decision Letter for (RSPB-2021-0906.R0)

See Appendix A.

Decision letter (RSPB-2021-0906.R1)

25-Jun-2021

Dear Miss Lonardo

I am pleased to inform you that your manuscript entitled "Dogs follow human misleading suggestions more often when the informant has a false belief" has been accepted for publication in Proceedings B.

You can expect to receive a proof of your article from our Production office in due course, please check your spam filter if you do not receive it. PLEASE NOTE: you will be given the exact page length of your paper which may be different from the estimation from Editorial and you may be asked to reduce your paper if it goes over the 10 page limit.

If you are likely to be away from e-mail contact please let us know. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

If you have any queries regarding the production of your final article or the publication date please contact procb_proofs@royalsociety.org

Data Accessibility section
Please remember to make any data sets live prior to publication, and update any links as needed when you receive a proof to check. It is good practice to also add data sets to your reference list.

Open Access
You are invited to opt for Open Access, making your freely available to all as soon as it is ready for publication under a CCBY licence. Our article processing charge for Open Access is £1700. Corresponding authors from member institutions (http://royalsocietypublishing.org/site/librarians/allmembers.xhtml) receive a 25% discount to these charges. For more information please visit http://royalsocietypublishing.org/open-access.

Your article has been estimated as being 10 pages long. Our Production Office will be able to confirm the exact length at proof stage.

Paper charges
An e-mail request for payment of any related charges will be sent out after proof stage (within approximately 2-6 weeks). The preferred payment method is by credit card; however, other payment options are available

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Thank you for your fine contribution. On behalf of the Editors of the Proceedings B, we look forward to your continued contributions to the Journal.

Sincerely,
Dr Robert Barton
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor:
Board Member
Comments to Author:
Thank you for responding thoroughly to all three reviewers' comments and for providing additional detail in your supplemental materials.

Appendix A

Associate Editor
Board Member: 1
Three reviewers have provided feedback on this article. All three highlight the importance of this article in expanding the focus of false belief understanding research beyond the study of primates. Furthermore, the reviewers acknowledge the value of pre-registration and the sound methodological design employed in this study. While all three reviewers are generally very favorable towards this study, all three provide detailed and thoughtful feedback as to how the authors can further clarify and enhance their reporting and I agree with the suggestions made. Furthermore, reviewer 2 requests a more critical appraisal of the results of experiment 3 and proposes alternative ways to consider the data more generally. I think such thoroughness would benefit the robustness of the conclusions drawn.

Thank you for considering our manuscript; we appreciate the positive feedback and the opportunity to submit a revision. Below we respond to the reviewers‘ comments and questions. We include a version of the manuscript where all changes were tracked. We believe that the comments led to a greatly improved version of the paper.

Referee: 1

Comments to the Author(s)
Through a novel cue-following object choice task, the authors provide the first evidence of dogs' sensitivity to others' beliefs. The study is of pronounced theoretical significance, and it is pre-registered, thoughtfully designed, unusually well-powered, clear and well-written. It includes a series of three studies within which the key results are clarified and replicated and an important alternative explanation is controlled for. Importantly, the authors do not attempt to over-interpret their results, clearly pointing (in the paper's second to last sentence) to several mechanistic hypotheses that can be tested in future research. The ability to track others' perspectives – and particularly others' beliefs – is at the heart of human sociality, from culture to cooperation to language. And consequently, for over forty years, researchers have been interested in whether this capacity is unique to humans. The authors provide the first evidence of potential false belief representation outside of primates. This paper is sure to garner substantial interest and inspire further research into the mechanisms and cognitive representations that support dogs' success in this task. It will also surely lead to investigations of the selective and environmental forces that produced convergent capacities in primates and dogs (and maybe other species, like corvids), and that resulted in divergent performance across breed classes. It is exceedingly rare to review a paper that is so strong and so obviously deserving of publication in nearly its current format. Well done!

We are grateful for the reviewer's very positive feedback, the constructive comments and time invested to improving our paper.

In my view, the authors have interpreted their results fairly in the discussion, including attempting to understand the unpredicted direction of their primary effects and pointing to alternative accounts for future research. The discussion could be further bolstered, however, by at least briefly addressing open questions about whether dogs' capacities are likely to be shared with wolves (or not) and whether they are likely to result from convergent evolution with primates (and maybe corvids) or shared ancestry. Relatedly, it could be valuable to highlight the need for future work that precisely characterizes the mechanisms across species, to determine the extent to which identical (as opposed to only superficially similar) mechanisms have evolved (perhaps convergently).

We agree with this suggestion and have expanded the discussion as follows: "the evolutionary origin of dogs' ability to distinguish between true and false beliefs of humans remains an open question. Future studies should examine how dogs and wolves (*Canis lupus*) compare in the current paradigm. If dogs' increased attention to human mental states results from the process of domestication, wolves are not likely to perform similarly to dogs in this task. Additionally, future research should clarify based on broader phylogenetic comparisons (e.g., comparing dogs with other domesticated species or primate species) whether identical or only superficially similar mechanisms have evolved across species and taxa" (lines 402-408). Based on this study (and species) alone, we preferred not to make hypotheses about possible evolutionary trajectories and about the likelihood of finding dogs' ability also in other species.

Minor points:

Line 39: this is true but the task was actually proposed by Dennett in his 1978 commentary "Beliefs about beliefs" in response to Premack and Woodruff's target article.

Line 64: note that Heyes' submentalizing hypothesis has been directly tested in apes:
Kano et al 2017 Submentalizing cannot explain belief-based action anticipation in apes. Trends Cogn Sci
Krupenye et al 2017 A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. Communicative & Integrative Biology

And in human infants:
Surian and Franchin 2020 On the domain specificity of the mechanisms underpinning spontaneous anticipatory looks in false-belief tasks. Dev Sci

All these papers are now mentioned in the manuscript.

Lines 176-215: when reading the procedure (i.e., before arriving to the results section), it was not clear which conditions were performed within each experiment. It would be helpful to make this explicit already in the procedure section for comprehension of the procedure and scoring and analysis sections.

We added this information in the Procedure paragraph (lines 209-211): "The dogs participating in Experiments 1 and 3, were presented either with the false belief (FB) or true belief (TB) event. Those participating in Experiment 2 were presented only with the control true belief (CTB) event."

Scoring and analysis: If space permits, I think it is worth moving the inter-rater reliability information to the main text

We could not move the inter-rater reliability information to the main text due to space limitations. The information is therefore reported in the ESM.

Procedure/scoring: I did not see information in the main text about (1) maximum trial length before it would be scored as no choice, or (2) whether trials were repeated in the event of no choice

Again due to space limitations, we had to leave these details in the ESM. However, we clarified there that if, after being released by their owner, dogs did not move for 30 seconds, we scored the trial as "no choice" and repeated it for a maximum of two times.

Lines 225-227: on first read, I misread the passage as indicting that all mentioned variables were included as random intercepts (which of course would not be possible). Separating fixed effects and random effects into separate sentences could ensure no confusion here.

Thank you for pointing this out to us, we separated the two sentences.

Scoring and analysis: The analysis strategy is sound but usually, when pursuing null hypothesis significance testing, a null model is created with only random effects and control variables (i.e., with test predictors removed) and null and full models are compared with a likelihood ratio test before the full model is interpreted. In the authors' case, their full models only include a single test predictor (experimental condition), meaning that the full-null comparison is exactly the same as the drop1 comparison used to generate the p-value for the test predictor. Maybe it is worth explicitly stating this for readers who will be wondering where the null model (and full-null model comparison) is?

We added this explanation to the Analysis section of the supplementary material. Please see also the answer below.

Analysis: If space permits, the authors should describe in the main text how significance of individual predictors was determined (drop1 function)

We added this information to the Analysis section of the supplementary material. The paragraph now states: "Because we were mainly interested in a single test

predictor (experimental condition), while the other fixed effects were considered control predictors, the comparison between the full model and null model lacking condition as predictor was implemented by using the R function drop1 (Chambers & Hastie, 1992) with argument 'test' set to "Chisq" to make inferences on the significance of the test predictor. The function drop1 drops each fixed effect from the model (one at a time) and uses a likelihood ratio test to compare the full with the respective reduced model (Barr et al., 2013). The pairwise comparisons were performed by re-levelling the reference category of condition. Wald tests were used to compare the performance of the control true belief condition (Experiment 2) to that of the true and false belief conditions of Experiment 1."

Analysis: The authors should also describe how pairwise comparisons were performed in cases when they did not derive directly from the models (e.g., in the model with all three conditions, were pairwise comparisons produced from re-levelling the variables or another function or were they based on other techniques distinct from the models themselves?

The pairwise comparisons were performed re-levelling the reference category of condition. We added this information to the Analysis section of the ESM.

Line 237: From the analysis section, I had the impression that Experiments 1 and 2 were analyzed in a single GLMM that included all 3 conditions. However, it becomes less clear in the results section whether all three conditions were analyzed together or whether the data from Experiment 1 were initially analyzed on their own. The supplement is clearer in this respect but this issue could be clarified in the main text.

We now clarified in the Analysis and Results sections of the main text that we initially compared only the true and false belief conditions of Experiment 1 using a first binomial GLMM (N=120). We subsequently compared the control true belief condition (Experiment 2) to the two conditions of Experiment 1 using a second binomial GLMM (N=180).

Referee: 2

Comments to the Author(s)
I really enjoyed reading this manuscript on human belief attribution by dogs. This is timely work and important results on belief attribution in a non-primate species. The results will very likely spark the general discussion on ToM-like skills in non-human species.

The authors tested dogs in two groups: one group observed a human pointing to a previously baited location, holding a true belief, while the other group observed a human holding a false belief. Using a preregistered study design for Experiment 1, they found that dogs in both groups differed in their behaviour – but in the opposite direction than the authors expected. After visual inspection of breed-specific variances in the choice data, the authors decided to run a follow-up study comparing border collies and terriers, where they could statistically support a difference in the group of terriers that is in line with their preregistered hypothesis. I am sincerely impressed by the comprehensive and detailed provision of behavioural data that made it very easy to follow the pre-defined analysis steps chosen by the authors (including open script, code, exploratory analysis and additional relevant statistics in the ESM).

We are very grateful for the reviewer's appreciation and thoughtful consideration of our work.

However, I have a few reservations regarding the interpretation of the data. One of my main concerns is about the rationale of Experiment 3 and the subsequent discussion of the different findings in Exp 1-2 and 3. I fully agree that breed-specific differences warrant further attention (and the authors themselves speculate in the discussion about a potential difference of more cooperative breeds vs more independent breeds). I am thus surprised that the authors only chose one specific FCI group to follow up on their initial results. In my opinion, this selection was premature and hampers, rather than advances, the discussion on the topic.

Before conducting Experiment 1, we did not expect to observe any breed difference in dogs' choices in this task and breed differences are not the main focus of the study. This explains why the subjects are not evenly distributed across FCI groups. With Experiment 3 we set out to follow up with new samples of dogs (in total, N=80) the behaviour observed in Experiment 1. Of course, given the large sample sizes required by the design, we had to limit ourselves in the number of FCI groups that we could additionally test (e.g., taking into account that we do not have access to an unlimited number of pure-bred dogs, that not all dogs were going to pass the familiarisation phase *etc.*). We chose to start by comparing terriers and border collies, but we do not believe that our choice will prevent future studies from targeting specifically the behaviour of other breeds (or groups of breeds). On the contrary, as stated in the discussion, future research on the phenomenon is needed. Given that this work was a first investigation of dogs' belief sensitivity, we did not deem necessary to provide already at this stage an exhaustive characterisation of

the behaviour in all FCI groups. Nevertheless, this should not hinder further investigation on the subject.

The authors state that terriers were the only ones that have chosen container A than B more often, while other groups showed no difference or the opposite pattern (286-287) – for me, it is not clear what criterion was used to make this inference (as e.g. FCI group 1 also shows a tendency to choose container A more often).

We reformulated more clearly in the text that we did not draw any inference but simply looked at the plot. In reply to this comment, we now better clarified our rationale and approach in the manuscript, lines 263-269: "Based on the visual inspection of dogs' choices in Experiment 1 (Fig. S2), afterwards further supported by an exploratory analysis (Fig. S8), the effect of experimental group (TB / FB) seemed to be completely reversed in one group (3: terriers), although no reliable conclusion could be drawn at that stage, given the small sample of terriers tested in Experiment 1 (N=10). To test the hypothesis that performance in this task might be influenced by the cooperativeness or independence of the breeds, we ran a follow-up experiment (Experiment 3) in which we tested a larger sample of terriers and border collies".

In addition, group 3 (terriers) only consisted of 10 subjects, and, if I am interpreting the wideness of the bars correctly, only 3 or 4 subjects were tested in the TB group – which makes it almost impossible to detect whether this pattern can be statistically supported.

In other words: as there were also other FCI groups with samples between 10-15 subjects, having 2-3 animals chosen differently in the TB group could have already made a difference in the visual inspection of the graph and the subsequent selection for the follow-up study.

Yes, by selecting terriers and border collies we did not mean to imply that other FCI groups or mixed breeds are undeserving of attention. Due to practical considerations concerning the feasibility of the study, we had to choose only a couple of breeds/group of breeds to begin with but otherwise extending the results to all existing breeds/mixes (and possibly to other canids) would have been very interesting and hopefully an endeavour for future research.

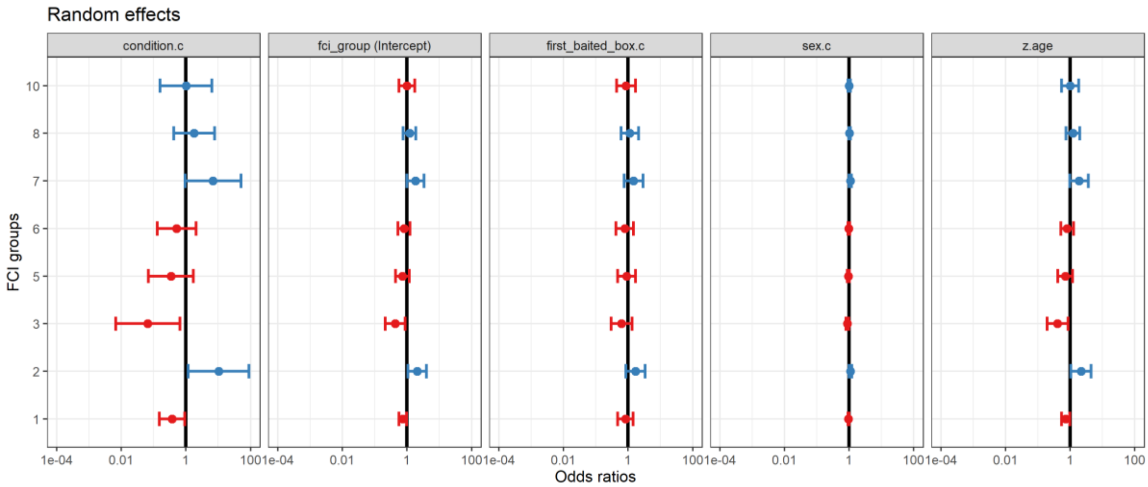
Given their data from Experiment 1 and their discussion on potential differences between more cooperative vs more independent breeds, why haven't the authors at least run an exploratory analysis from their initial dataset to potentially support their alternative explanation?

After Experiment 1, we did not know whether the variance observed within the different FCI groups was random or due to a real difference in the population. As correctly noted above, the data of Experiment 1 was not analysed a second time, subsetting subjects according to their FCI group, due to the varying and sometimes

small sample sizes across groups that could have undermined the reliability of statistical findings.

We decided which breeds to include in the follow up based on visual inspection of the plots and the likely availability of dogs. In short, instead of exploring the data of Experiment 1 statistically, we decided to test the hypothesis that there might be a difference between the behaviour of cooperative and independent workers directly with new samples of dogs in Experiment 3.

However, in response to these comments, we fitted a binomial GLMM to the data of all dogs tested in Experiment 1. All fixed effects were as in the other models but we included FCI group (instead of breed) as random effect (code available in the GitHub repository). We found that terriers' choices were the only ones whose confidence interval does not overlap with an odds ratio of 1:



We added to the ESM Figure S8 (reported also here above) with the following caption: "Random effects of a binomial GLMM with FCI group as random intercept (N=144, all dogs in Experiment 1). The same predictor variables as in GLMM01 were included (condition, first baited box, sex, age) in the model. The only difference was that FCI group was included as random intercept instead of breed (as well as all possible random slope components within FCI group). This analysis was performed, after seeing the results of Experiment 3, in support of the visual inspection that led to the follow-up test of FCI group 3 (terriers) in Experiment 3. As shown in the first panel "condition.c", FCI group 3 (terriers) is the only one whose confidence interval does not overlap with an odds ratio of 1."

I was additionally wondering why the authors in Exp3 have not chosen a range of breeds from FCI 1 and opted for one single breed instead – in particular, as the authors did not focus on one breed of terriers.

We would have preferred to only test one breed from FCI group 3 as well, to reduce variation in the sample. However, because owners of pure-bred terriers are less common in the population of dog owners who volunteer to participate in our experiments, we had to recruit different terrier breeds in order to obtain a final sample size of 40 individuals. Instead, within FCI group 1, border collies were the

most popular breed in our database and we knew we could have reached the predetermined sample size with just one breed.

I wanted to emphasize that the authors might consider also to discuss that a sample of dogs that is prone to the perseveration bias (A not B error) has likely been excluded from the tested population (following Fam phase 1).

Yes, because the exclusion criterion was not too loose, it allowed us to filter such dogs. In general, we chose not to devote so much attention to the excluded dogs as their performance is not informative about the main topic of the paper. However, given the interest that they raised, we reported more information in addressing the comments below.

As performance on this task could potentially be linked to difference in attention (to the task) and/or gathering information based on local enhancement, it would be interesting to know how many dogs from each FCI group had to be excluded from the test based on what criterion.

Of the 52 dogs excluded from Experiment 1 because they did not reach the inclusion criteria, one dog (Hungarian short-haired pointer) refused to enter the test room prior to the first familiarisation trial. Eighteen dogs failed to make two correct choices in two consecutive trials (one without and one with displacement) within 4 trials during the first familiarisation phase. Four dogs (two from FCI group 8, one from FCI group 5 and one from FCI group 3) did not make a choice (i.e., they did not leave the owner's side for at least 30 seconds from when they were released, despite the owner's verbal encouragement) in two consecutive trials during the first familiarisation phase.
Given the nature of the question, we report in the table below (column "Exp. 1 Fam. 1") only the eighteen dogs that were excluded for making mistakes.
Twenty-one dogs were excluded from Experiment 1 because they did not reach the inclusion criterion of phase 2 and eight because they did not reach the inclusion criterion of phase of phase 3.

Of the 24 dogs excluded from Experiment 2, four (two dogs from FCI group 10, one from FCI group 7 and one from FCI group 1) did not make a choice in more than two consecutive trials during the first familiarisation phase.
Again, we report in the table below (column "Exp. 2 Fam. 1") only the 10 dogs that were excluded for failure to reach the inclusion criterion of phase 1.
Seven dogs were excluded from Experiment 2 because they did not reach the inclusion criterion of phase 2 and three because they did not reach the inclusion criterion of phase of phase 3.

Of the 22 dogs excluded from Experiment 3, one (Jack Russel terrier) was excluded due to two consecutive "no choice" trials during the first familiarisation phase and one (West Highland white terrier) due to two consecutive "no choice" trials during the second familiarisation phase. Finally, one (Parson Russel terrier) became too

agitated to continue with the experiment in phase 3. Once more, we report in the table below only the 19 remaining dogs, that were excluded because they did not reach the inclusion criteria of phases 1 (eleven dogs), 2 (six dogs) and 3 (two dogs). We added to the ESM Table S10, reported also here below for convenience.

FCI group	**Exp. 1 Fam. 1**	**Exp. 1 Fam. 2**	**Exp. 1 Fam. 3**	**Exp. 2 Fam. 1**	**Exp. 2 Fam. 2**	**Exp. 2 Fam. 3**	**Exp. 3 Fam. 1**	**Exp. 3 Fam. 2**	**Exp. 3 Fam. 3**
1	**4**	**7**	**3**	**4**	**2**	**2**	**5**	**4**	**1**
2	**1**	**1**	**1**	**3**	**2**	**0**			
3	**4**	**2**	**1**	**0**	**0**	**0**	**6**	**2**	**1**
5	**1**	**2**	**0**	**0**	**0**	**0**			
6	**1**	**0**	**0**	**0**	**0**	**0**			
7	**0**	**1**	**0**	**0**	**2**	**1**			
8	**5**	**6**	**2**	**2**	**1**	**0**			
9	**1**								
10	**1**	**2**	**1**	**1**	**0**	**0**			

On a last general remark, I think that the introduction and discussion are in parts quite lengthy, while other parts need some additional crucial information to provide the reader with a better walk-through of the study rationale (see detailed comments).

In response to this comment, we deleted the parts of the introduction that were redundant with the methods section and we added to the paper all the information required by the reviewer (details in the replies below)

Detailed comments:

General
Please provide an ethical approval number

We have now included it in the first paragraph of the Methods section.

Intro

Why would we expect differences in a cooperative and a competitive task? How does a competitive task look like? Given that the authors have chosen a cooperative task, this appears like crucial background to understand the study rationale.

Despite the popularity of speculations on how the cooperative/competitive nature of tasks might influence apes and dogs' performance (e.g., Hare et al., 2001; Hare & Tomasello, 2005; Krachun et al., 2009) so far, these hypotheses do not seem to find support in empirical evidence. Indeed, the only explicit false belief test that yielded

positive results with apes involved a helping paradigm in which participants helped the experimenter obtain an object of interest. And evidence for dogs' capability of perspective taking comes from both cooperative (e.g., Catala et al., 2017) and more "agonistic" (Heberlein et al., 2017; Kaminski et al., 2013) situations.
When engaging with conspecifics in social play that involves objects, dogs tend to be more competitive than when playing with humans (Rooney et al., 2000). In competitive situations (involving dog and human) dogs might display conflict-minimising behaviours, that might potentially obscure the ability of interest. Hence, we chose to embed the false belief task in a (mostly) cooperative situation. Also, previous studies suggest that dogs excel at reading human communicative intentions and might interpret some repeated hiding-finding interactions as a social game with the experimenter (e.g., Topal et al., 2009; Topál et al., 2005).
Due to space limitations, we included a reduced version of this answer in the Introduction (lines 84-87).

38 a little bit more background on the importance of the seminal article would be advantageous for readers unfamiliar with the field

Following Reviewer 1's suggestion, we now mentioned in the Introduction also Dennett's commentary to the seminal article. Due to space limitations, however, we have to refer the reader interested in the philosophical and empirical work behind the change of location paradigm to the original literature on the topic.

40 a bracket is missing at the end of the sentence

48 please add the Latin name for chimps

We have resolved the two issues above.

58-64 I think it would add to the comprehension of these criticisms if the authors would suggest a potential solution to them

We now included in the Introduction the studies with apes and infants that used the procedures suggested by Heyes to solve these criticisms. These studies controlled for alternative explanations such as the behavioural rule and submentalizing accounts (Kano et al., 2017, 2019; Krupenye et al., 2017; Surian & Franchin, 2020).

66 here the authors mention that dogs are a particularly interesting case because of their shared social environment – as this would likely also be the case for the common housefly, I think the authors should elaborate why dogs are good human behaviour readers (and might constitute a different case than, e.g., other domestic animal species)

Indeed, the performance of other domestic species in this task would be interesting as well. However, we expanded the introduction by explaining in particular how dogs' socio-cognitive capabilities and their understanding of human

communication make them an exceptionally suitable model for the comparative study of human social cognition.

74-109 I think a lot of the information covered here should actually be placed in the method section (e.g. 79-82; 89-92)

We removed part of the indicated text from the Introduction as it was indeed a repetition of what was written in the Methods and we moved the rest to the Methods.

114 the authors should elaborate here what they mean with retroactive interference

We moved at this point the definition of retroactive interference (lines 116-117).

120-137 I think that some details here could be cut down (and moved to the discussion)

We moved this part to the discussion.

137 Reference(s) missing

We added a few references (Chapagain et al., 2017; Udell et al., 2014; Karl et al., 2020) in line 386.

Methods

Familiarisation phase 3: "Only dogs that made two correct choices in two consecutive trials (one with displacement, one without displacement) within four trials were subsequently tested in the final test phase." Why has this rather weak criterion been chosen?

The exclusion rate (98 out of 260 dogs did not reach the test because of this criterion) strongly suggests that the criteria were not too loose and a more conservative choice would have probably made the final sample too biased and unlikely to represent the general population.
However, even before conducting the experiments, we had chosen not to have a training phase. Rather, we opted for a quick familiarisation that was necessary to ensure that we only tested dogs that payed attention to the events, understood that only one piece of food was hidden in each trial, were motivated to find the reward and were comfortable in the laboratory setting. The reason we avoided to have a more demanding inclusion criterion was that we aimed at investigating pet dogs' *spontaneous* belief sensitivity (i.e., in the absence of previous training).

What posthoc test was used for the analysis of Exp1-2?

A Wald test was used to compare the performance of the control true belief condition (Experiment 2) to that of the true and false conditions of Experiment 1.

We included this information in the ESM.

151-152 please provide more details – why match with TB, but not FB group?

In reality, there is no difference between matching dogs in the CTB group with those in the true belief group or with those in the false belief group because the dogs in the two groups of Experiment 1 were also matched at the breed level as much as possible. However, because for the counterbalancing we also considered sex, age and first baited bucket, and because we did not know in advance which dog breeds we would have been able to recruit and test, slight variations in the breeds characterise the two groups of Experiment 1 (see the Table S9, reported also here below for convenience).
When choosing which one of the two groups we needed to match more closely in terms of subjects' breed for Experiment 2, to test the retroactive interference hypothesis, we chose the true belief one. The reasoning was as follows: dogs in the false belief group witnessed an additional possibly distracting event (the communicator re-entering the room) after the final hiding of food in bucket B. This event might explain why more dogs chose the empty bucket in the false belief group than in the true belief group (where dogs did not witness this event *after* the hiding of food in bucket B but before). Therefore, if retroactive interference, and not the communicator's belief, explains dogs' behaviour in the FB group, also dogs in a true belief condition, if they witnessed the possibly distracting event after the displacement of food, should react in the same way as those in the FB group. If, however, the belief of the communicator explains dogs' behaviour in the false belief group, dogs in a true belief condition should continue to ignore the cue more often than in the false belief group irrespective of the moment of re-entry of the communicator.
Hence, in short, we wanted to find out whether dogs in the control true belief group would have still been able to ignore the misleading cue at a similar rate as dogs in the original true belief group despite witnessing even more possibly distracting events (in addition to the misleading suggestion) than dogs in the false belief group. Because this is what we found, we concluded that retroactive interference is an unlikely explanation for the behaviour of dogs in the false belief group too. This reasoning is summarised in lines 240-253 of the manuscript.

We included in the ESM Table S9, showing the number of dogs in each FCI group and treatment across Experiments 1 and 2 (N=180).

FCI group	**False Belief**	**True Belief**	**True Belief Control**
1	**19**	**21**	**21**
2	**7**	**7**	**7**
3	**4**	**1**	**1**
5	**2**	**4**	**4**
6	**5**	**8**	**7**
7	**7**	**7**	**6**
8	**8**	**10**	**11**

10	**6**	**4**	**3**

Discussion

282-292 Summarising the rationale in the first paragraph of the discussion is a great way to remind the reader about the key questions that the authors aimed to answer – I would additionally favour that the authors would also include a very brief summary of the results (and the inferences that they can draw from them) in this paragraph

We agree and we included the indicated information in lines 281-285: "This study aimed at investigating whether dogs would spontaneously behave in a different way in response to a misleading suggestion from a human informant with a true (TB) or a false belief (FB) and this is indeed what we found in Experiment 1. The combined results of Experiments 1 and 2 suggest that retroactive interference is not a likely explanation for the behaviour of dogs in this task. Finally, the results of Experiment 3 show that performance in this task is subject to breed (group) differences".

293-310 I think these two paragraphs should be merged. I also think that prior to discussing the control condition from Exp2, the results of Exp1 should be discussed in a bit more detail

We merged the two paragraphs. The results of Experiment 1 are discussed more in detail after the results of Experiment 2 because we wanted to make clear for the reader that the results of Experiment 1 are (likely) not due to retroactive interference or distraction before discussing them further.

311-12 the authors should elaborate here why this result was surprising, e.g. by stating the directionality of their hypothesis prior to this statement

We agree, and reiterated at this point our initial prediction.

331-333: even if dogs perceive the misleading pointing in the FB group as a mistake in goodwill (or any mistake for whatever reason), they should still see it as a mistake and choose the baited container, right? Or are the authors aiming to make the case that the human in the FB group is still be seen as a collaborator (rather than the one in the TB group -> deceiver) and that this difference in perceived trustworthiness gives dogs a higher inclination to be misled? What would be the mechanisms that would cause such an inclination?

Based on the literature on dog-human social interactions and on this study, we are not sure that upon realizing that a human is mistaken, dogs would behave in the "correct" manner (i.e., in this case, choose container B). Indeed, as we had stated in lines 338-343, dogs' social bias, the tendency to make counterproductive choices

under the influence of a human demonstrator, is a well-documented phenomenon (e.g., Barnard et al., 2019; Kupán et al., 2011; Marshall-Pescini et al., 2011, p., 2012; Prato-Previde et al., 2008; Szetei et al., 2003). We agree with the authors of these studies in the interpretation of the behaviour: it might be that, from a dog's point of view, maintaining or enhancing social cohesion with humans (which is done by not contradicting/correcting the mistaken human) is more important than obtaining the food.
Additionally, two studies using different paradigms (Kubinyi et al., 2003; Topál et al., 2005) suggest that dogs might understand certain repetitive situations as a social game between them and a human (experimenter/owner). For example, dogs in a hiding-finding game kept searching at old locations a toy they knew was no longer hidden there. The authors suggest that this behaviour should be interpreted as "rule-following"; i.e., during the warm-up trials dogs came to interpret the experiment as a game based on the rule: "the experimenter hides, I search". The authors speculate that such a rule-following behaviour might minimise social conflicts with humans and consider rule-following as one of the aspects of the wider phenomenon of dog-human "synchronisation" (Miklósi & Topál, 2012; Topál et al., 2009). Hence, the communicator with a false belief might be perceived as a trustworthy, albeit mistaken, informant who is still playing the same game as in the familiarisation, while the communicator with a true belief (suddenly switching to uncooperativeness) might be perceived as less trustworthy or violating the rule of the game. We added part of these considerations to the discussion (lines 345-353). However, based on the current state of knowledge, it is not possible to answer this question conclusively. Considering also that we initially expected the opposite pattern of results relative to the one we found, we did not deem it appropriate to further speculate in the manuscript on the mechanisms underlying dogs' behaviour. Other studies, designed to target specifically the proximate causes, are needed.

348 I think it would be very informative to also have the number of subjects stated that were assigned to each treatment group from each FCI group (e.g., in the ESM). Given the difference in numbers between FCI groups (and FCI groups x treatment groups) it might be better to change phrases such as "more terriers" to "relatively more terriers" etc.

Following the reviewer's recommendation, we added Table S9 to the ESM. All 4 terriers in the TB group of Experiment 1 chose container A, while only 1 terrier (out of 6) in the FB group chose the empty container. Therefore, the expression "more terriers" is correct at this point (line 360). However, we added "*relatively* more terriers" in line 430, where we compare the performance across FCI groups with largely differing sample sizes.

362-365 Why would this specifically be the case for terriers (see also the main comment on why terriers have been chosen for exp3)?

We do not know whether this might be specifically true for terriers. On the contrary, in our speculation we generalise quite broadly to cooperative and independent workers, based on the 2 instances we tested (border collies and

terriers). Our post-hoc hypothesis needs to be tested in future studies with other samples of cooperative and independent workers (hence, all FCI groups that we had to ignore due to practical constraints), as we stated in the discussion.
As stated also above, terriers were just one of the possible groups that we could have chosen. We do believe that investigating the performance of all dog breeds in this task would be a valuable addition to the literature. However, due to practical constraints on time and resources, we had to limit ourselves in the number of breeds we tested. Faced with this choice, we went for terriers because the 10 terriers we tested in Experiment 1 clearly showed the opposite pattern of response (a fact, directly visible from Figure S2 and now confirmed by a random effect analysis) and because we were confident that we might have found at least 40 terriers to include in the sample. This would have not necessarily been the case for other independent workers. Because we knew it would have not been possible to find more than 40 terriers belonging to just one specific breed, we had to recruit multiple terrier breeds, although in an ideal world we would have preferred to reduce the variation in the sample by recruiting just one breed.

ESM

"For Experiment 3, we conducted another power simulation to determine the sample size. This revealed a power of 76% with 40 dogs and an expected performance of 0.3 in the FB group and 0.7 in the TB group (performance predictions based on the terriers performance in Experiment 1).". As the performance of terriers was about 0.2 in the FB group and 1.0 in the TB group, I assume that the authors choose more conservative numbers based on the low sample size in Exp1?

Yes, exactly. We clarified this point in the ESM.

Did the authors observe breed differences in anticipatory looking? I am just wondering what were the reasons the authors opted to not include it (N is likely to low?), as all FCI group data has been plotted for choice and latency, but not anticipatory behaviour

We agree that for exhaustiveness of reporting, also this variable should be plotted. We had not included it initially due to the very small sample sizes in most of the FCI groups, which undermine the reliability of any conclusion. However, we added to the ESM Figure S7, showing, for each FCI group, the proportion of dogs that looked at container B before the communicator's suggestion in the TB and FB condition of Experiment 1. We also report in the plot, under each bar, the number of dogs in that condition and FCI group that gazed at least at one container (A or B) before the communicator's suggestion.

Plot width differs in e.g. Figure ESM S2 – I cannot find information in the legend, but might assume that the bar widths represents the sample size for each treatment (similar to the other figures)?

Yes, correct. Thank you for pointing this out: we added the information to the legend of Figure S2.

Referee: 3

Comments to the Author(s)
Review on the manuscript
Dogs follow human misleading suggestions more often when the informant has a false belief
Written by Lonardo et al.

Overall opinion
This is an interesting and well-written study, where authors experimentally tested whether dogs act differently in a visible food displacement two-choice task when they are supported by either a knowledgeable or non-knowledgeable human informant (the 'communicator'). As the informant always suggested to the dogs to choose the 'wrong' container (where the food was moved from to the other container), the information can be regarded as false. Thus, dogs' eventual choice patterns can be regarded as whether they were dependent on the 'false belief' on the non-knowledgeable informant.
The experiments are carefully designed and analyzed, they were run on an impressive size sample. The results are interesting and their interpretation is modest. Authors carefully took care of the important alternative hypotheses.

We thank the reviewer for this positive evaluation of our work as well as for the attentive feedback.

My critical comments are mostly minor (listed below). I found two issues that can represent some level of concern. One is the inclusion of rather young (5-12 months of age) subjects to the sample. The other is a somewhat tricky question: can we be sure that dogs would chose the 'good' container without any intervention of the 'communicator'? Authors should provide a good reasoning for this, unless they do not want to include one more control group (which I doubt, understandably).

Please find detailed responses below.

Detailed comments
I would recommend to use an important piece of literature that is one of the rare attempts in the past to test dogs' performance in a task where a human 'helper' had correct or incorrect knowledge about the task (hence, this paper can be regarded as testing for 'false belief' attribution in dogs):
Virányi, Z., Topál, J., Miklósi, Á., & Csányi, V. (2006). A nonverbal test of knowledge attribution: a comparative study on dogs and children. Animal Cognition, 9(1), 13-26.

We included the study in the introduction as further evidence in favour of dogs' ability to take into account humans' past perceptual access.

The manuscript has a very odd, somewhat confusing structure. After the true Introduction comes a long section (lines 74-137), which is formally still within the Intro, but in reality it is a mixture of methods and hypotheses. I do believe that the research goal, questions, predictions and hypotheses should be placed to the end of the Introduction, but in this case it was done in a too large extent. Especially the detailed methodology seems like a repetition, because later in the Methods this is done again.

We agree and removed the description of the methodology from the Introduction.

Could you provide a justification for the inclusion of both juvenile (5-12 months) and adult dogs to the test population? Although there are results that show juvenile (or even younger) dogs performing comparably well in tasks that involve inter-specific social cognition, there can be factors that affect differently the performance of juvenile and adult dogs (level of training, attention span, inhibitory threshold etc.).

The number of dogs tested between 5 and 11 months of age was: 6 in Experiment 1, 7 in Experiment 2 and 10 in Experiment 3 (5 border collies and 5 terriers).
We believe that the exclusion (more than the inclusion) of juvenile dogs from this study would have needed justification.
First of all, Barnard et al., (2019) showed that the tendency to conform to human misleading suggestions is present in puppies already at 4 months. In response to this comment, we included only this information in the Discussion (lines 343-345) due to space constraints.
However, as the reviewer correctly points out, other studies (e.g., Bray et al., 2021; Hare et al., 2002; Riedel et al., 2008) showed that already from 6-8 weeks of age puppies are capable of following human communicative gestures to locate hidden food in a cooperative context.
Second, we had three familiarisation phases to ensure that all the tested dogs (irrespective of their age) behaved in a way that was compatible with their "understanding" of the game (i.e., they payed attention to the movements of the experimenters and food, they chose first the baited bucket despite the displacement, etc.).
Third, the finding that younger dogs were more likely to follow the communicator's cue irrespective of condition reveals an interesting similarity with humans. This similarity could shed some light on the mechanism underlying dogs' sensitivity to a human misleading influence in a situation where dogs' knowledge conflicts with the informant's suggestion. Hence, an interesting similarity with humans might have been missed had we not included juvenile dogs in the study.

How long was the pause between baiting bucket A and transferring the food to bucket B in case of the three initial familiarization trials? Was it comparably long to the pause that they used in the testing trials, or was it shorter?

During the displacement trials of the familiarisation phases, the food remained in bucket A for less than 2 seconds before being moved to bucket B. The amount of time between the initial baiting of bucket A and the displacement of the food was longer during the test trials because the communicator had to leave the room and, in the true belief condition, also re-enter, before the hider could move the food (the communicator never left the room during the familiarisation). Instead, the

same amount of time elapsed between the initial baiting of bucket A and the displacement of the food during the familiarisation and the TBC test condition. However, we were not interested in comparing the performance of dogs between the familiarisation and the test trials.
In response to this comment, we provided in the ESM (paragraph "Timing of the food displacements") the following information about the durations of the test events: "In the True Belief (TB) test trials, the food remained in bucket A for approximately 42 seconds before being moved to bucket B. In the False Belief (FB) test trials, the food remained in bucket A for approximately 12 seconds before being moved to bucket B. In the Control True Belief (CTB) test trials, the food remained in bucket A for less than 2 seconds, as during the familiarisation trials with displacement. In both the TB and FB trials, the communicator stayed outside of the room for approximately 20 seconds and in the CTB trials, the communicator stayed outside of the test room approximately 1 second."

What sort of food reward was used for baiting? Were the pieces large enough for the dogs to see their transfer to and from the buckets?

We added to the ESM that we used for the majority of dogs dry food pellets (ca. 1 cm thick) and for dogs who were not interested at all in dry food (as judged by the hider upon arrival of the dog to the lab) slices of sausage (ca. 1 cm thick).
In case of food allergies or special dietary requirements, owners brought to the lab their dog's usual food. When needed, this was cut into pieces approximately 1 cm in diameter.
As stated in the Methods, we used visible displacements. The hider took care to show the piece of food (and not simply her hand holding the food) to the dog every time the food was moved. The hider would always move as close to the dog as necessary for the dog to see the food. Additionally, every time food was hidden in a bucket, dogs could hear the sound of the piece of food falling into the bucket. The only time dogs could not see the food being sneakily removed from the bucket (at the end of the test trials) this happened because the hider hid the piece of food in her closed fist, instead of openly showing the piece of food to the dog.

Did dogs always choose the bucket that was suggested by the communicator in the warm-up trials?

No, not always (meaning not all dogs, indeed some were excluded from the study). The number of dogs that were excluded from familiarisation phases 2 and 3 (the ones with communicator) for choosing the "wrong" bucket is now documented in the supplementary materials (Table S 10). However, the communicator's cue seemed effective (see ESM) as 140 out of the 144 dogs tested in Experiment 1 followed this cue the first time they saw it.

Line 269 – please add the names of the particular FCI dog breed groups here to the results. Referring to their FCI-numbers only, hides the interesting nature of the results, which shows remarkable pattern according to the breed groups.

We reformulated this passage in accordance with Reviewer's 2 comments, hence the FCI groups are no longer mentioned at this point. Instead, we only talked about the difference between cooperative and independent workers.

Line 328 – Authors here state that "Along this line of argument, dogs in both groups remembered the final location of food (bucket B)." Actually, we cannot know this, because there was no control group that would test dogs' choices WITHOUT the interference of the 'communicator' experimenter.
In such a control group, everything would happen as in the False Belief group, but the 'communicator' would not suggest any of the buckets for the dogs before those are let to make their choice.

We agree that the communicator suggesting the empty bucket A is a potentially interfering event: something that could make dogs' mnemonic trace (and hence their choice of bucket B) decline rather than increase. Hence, removing the communicator's suggestion (somehow similarly to the events of the familiarisation phase 1) should have no influence on dogs' performance, or it could even improve it. Instead, we cannot think of a reason why removing a possible source of interference should degrade the performance. Indeed, previous studies using a similar situation, a misleading human suggestion in a quantity discrimination task (e.g., Marshall-Pescini et al., 2012), found that with no interference from the demonstrator (No Influence condition), dogs showed an overall preference for the larger food quantity and only started choosing against this preference under the influence of the human demonstrator.

Furthermore, if dogs in our study had no memory of which container was baited last in the test trial, similarly to the second familiarisation phase, in which they could not know where food had been hidden in their absence, they should have followed the communicator's suggestion (and hence chosen bucket A) at a much higher rate than they did in all three conditions. Indeed, as reported in the ESM, 140 out of 144 dogs followed the communicator´s cue in the first trial of the second familiarisation phase (binomial test; p-value < 0.001) when they had not witnessed themselves the hiding of food.

However we can see that the wording of ours was too strong and changed it accordingly to: "along this line of argument, *it seems plausible to assume that* dogs in both groups remembered the final location of food (bucket B)".

References

Barnard, S., Passalacqua, C., Pelosi, A., Valsecchi, P., & Prato-Previde, E. (2019). Effects of breed group and development on dogs' willingness to follow a human misleading advice. *Animal Cognition*, *22*(5), 757–768.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bray, E. E., Gnanadesikan, G. E., Horschler, D. J., Levy, K. M., Kennedy, B. S., Famula, T. R., & MacLean, E. L. (2021). Early-emerging and highly heritable sensitivity to human communication in dogs. *Current Biology*.

Catala, A., Mang, B., Wallis, L., & Huber, L. (2017). Dogs demonstrate perspective taking based on geometrical gaze following in a Guesser–Knower task. *Animal Cognition*, *20*(4), 581–589. https://doi.org/10.1007/s10071-017-1082-x

Chambers, J. M., & Hastie, T. J. (1992). Linear models. Chapter 4 of statistical models in S. *Wadsworth & Brooks/Cole*.

Hare, B., Brown, M., Williamson, C., & Tomasello, M. (2002). The domestication of social cognition in dogs. *Science*, *298*(5598), 1634–1636.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, *61*(1), 139–151.

Hare, B., & Tomasello, M. (2005). Human-like social skills in dogs? *Trends in Cognitive Sciences*, *9*(9), 439–444.

Heberlein, M. T. E., Turner, D. C., & Manser, M. B. (2017). Dogs' (Canis familiaris) attention to human perception: Influence of breed groups and life experiences. *Journal of Comparative Psychology*, *131*(1), 19–29. https://doi.org/10.1037/com0000050

Kaminski, J., Pitsch, A., & Tomasello, M. (2013). Dogs steal in the dark. *Animal Cognition*, *16*(3), 385–394.

Kano, F., Krupenye, C., Hirata, S., Call, J., & Tomasello, M. (2017). Submentalizing Cannot Explain Belief-Based Action Anticipation in Apes. *Trends in Cognitive Sciences*, *21*(9), 633–634. https://doi.org/10.1016/j.tics.2017.06.011

Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the National Academy of Sciences*, *116*(42), 20904–20909. https://doi.org/10.1073/pnas.1910095116

Krachun, C., Carpenter, M., Call, J., & Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science*, *12*(4), 521–535.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2017). A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. *Communicative & Integrative Biology*, *10*(4), e1343771. https://doi.org/10.1080/19420889.2017.1343771

Kubinyi, E., Miklósi, Á., Topál, J., & Csányi, V. (2003). Social mimetic behaviour and social anticipation in dogs: Preliminary results. *Animal Cognition*, *6*(1), 57–63.

Kupán, K., Miklósi, Á., Gergely, G., & Topál, J. (2011). Why do dogs (Canis familiaris) select the empty container in an observational learning task? *Animal Cognition*, *14*(2), 259–268. https://doi.org/10.1007/s10071-010-0359-0

Marshall-Pescini, S., Passalacqua, C., Miletto Petrazzini, M. E., Valsecchi, P., & Prato-Previde, E. (2012). Do Dogs (Canis lupus familiaris) Make Counterproductive Choices Because They Are Sensitive to Human Ostensive Cues? *PLoS ONE*, *7*(4), e35437. https://doi.org/10.1371/journal.pone.0035437

Marshall-Pescini, S., Prato-Previde, E., & Valsecchi, P. (2011). Are dogs (Canis familiaris) misled more by their owners than by strangers in a food choice task? *Animal Cognition*, *14*(1), 137–142.

Miklósi, Á., & Topál, J. (2012). The evolution of canine cognition. *Oxford Library of Psychology. The Oxford Handbook of Comparative Evolutionary Psychology*, 194–213.

Prato-Previde, E., Marshall-Pescini, S., & Valsecchi, P. (2008). Is your choice my choice? The owners' effect on pet dogs'(Canis lupus familiaris) performance in a food choice task. *Animal Cognition*, *11*(1), 167–174.

Riedel, J., Schumann, K., Kaminski, J., Call, J., & Tomasello, M. (2008). The early ontogeny of human–dog communication. *Animal Behaviour*, *75*(3), 1003–1014.

Rooney, N. J., Bradshaw, J. W., & Robinson, I. H. (2000). A comparison of dog–dog and dog–human play behaviour. *Applied Animal Behaviour Science*, *66*(3), 235–248.

Surian, L., & Franchin, L. (2020). On the domain specificity of the mechanisms underpinning spontaneous anticipatory looks in false-belief tasks. *Developmental Science*, *23*(6). https://doi.org/10.1111/desc.12955

Szetei, V., Miklósi, Á., Topál, J., & Csányi, V. (2003). When dogs seem to lose their nose: An investigation on the use of visual and olfactory cues in communicative context between dog and owner. *Applied Animal Behaviour Science*, *83*(2), 141–152. https://doi.org/10.1016/S0168-1591(03)00114-X

Topál, J., Gergely, G., Erdohegyi, A., Csibra, G., & Miklosi, A. (2009). Differential Sensitivity to Human Communication in Dogs, Wolves, and Human Infants. *Science*, *325*(5945), 1269–1272. https://doi.org/10.1126/science.1176960

Topál, J., Kubinyi, E., Gácsi, M., & Miklósi, Á. (2005). Obeying social rules: A comparative study on dogs and humans. *Journal of Cultural and Evolutionary Psychology*, *3*(3–4), 223–243.

Topál, J., Miklósi, Á., Gácsi, M., Dóka, A., Pongrácz, P., Kubinyi, E., Virányi, Z., & Csányi, V. (2009). Chapter 3 The Dog as a Model for Understanding Human Social Behavior. In *Advances in the Study of Behavior* (Vol. 39, pp. 71–116). Elsevier. https://doi.org/10.1016/S0065-3454(09)39003-8