natureresearch

Peer Review Information

Journal: Nature Human Behaviour
Manuscript Title: Paranoia and belief updating during the COVID-19 crisis
Corresponding author name(s): Philip R. Corlett

Editorial Notes:

Reviewer Comments & Decisions:

Decision Letter, initial version:

18th March 2021

Dear Dr Corlett,

Thank you once again for your manuscript, entitled "Paranoia and belief updating during a crisis". I apologize for the delay in the peer review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of considerable potential interest, they have also raised quite serious concerns. In light of these comments, we cannot accept the manuscript for publication in its current form. We would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

In particular, as well as addressing all of the other reviewer points, it will be very important to respond to the questions raised by Referee #2 and #3 regarding preregistration (or lack of it). You should clearly state whether (and which of) your analyses were preregistered; if the work wasn't preregistered, you should clearly indicate what was hypothesized a priori and why (refraining from hypothesizing after the fact), and which analyses were exploratory. Whether the analyses were preregistere or not, you will also need to correct for multiple comparisons as Referee #3 suggests. We will

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript we would hope to receive it within 8 weeks (2 months). If you are unable to submit your revised manuscript within 8 weeks, please let us know.

With your revision, please:

• Include a “Response to the editors and reviewers” document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

<strong>Note:</strong> This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,
Jamie

Dr Jamie Horder
Senior Editor
Nature Human Behaviour

REVIEWER COMMENTS:

Reviewer #1:
Remarks to the Author:
I would like to recommend this article for publication.

This is an excellent article probing paranoia-relevant cognition in the US population before and during the COVID-19 pandemic. The article identifies specific cognitive features revealed by the computational model, which differ between high and low paranoia groups examined. These cognitive features, such as the expected reward rate that paranoid individuals have (higher than controls) manifest themselves in both the versions of the reversal task that the authors administered. In addition, the authors examined a small number of public health measures that varied across the US states whence the participants came from, especially whether states mandated or recommended the use of masks, what the levels of compliance to this measure were, the propensity of states to tightly follow rules, and how these variables related to paranoia.

I think the demographic data is interesting, the analyses are adequate, and various recommendations that could be made in theory (such as the usage of longitudinal data, establishing the state vs. trait 'paranoia' involved etc., comprehensively assessing mental health symptoms and establishing rigorously the validity of the claim 'significantly. Anxiety increased ... but, the change was less pronounced ... than paranoia (Figure 2a), suggesting a particular impact of the pandemic on beliefs about others.') would be out of place in the context of a pandemic. However I would like to point out that using unemployment as an indicator of socioeconomic distress is way too crude a measure to examine the impact of socioeconomic factors here. There are numerous much better measures of socioeconomic effects.

However there are two points where I strongly believe that the article fails, which need to be corrected.
The first is in the portrayal of the hierarchical gaussian filter generative model. The paragraph 'Our generative model, the hierarchical ...' is very poor for readers who are not familiar with the HGF. It is way too brief an introduction, the reader having no picture of what the HGF really is, and why it is relevant to paranoia. Terms like 'tonic uncertainty' are not defined. It is sloppy with the terminology - how can a mean be a 'prior belief' - surely prior beliefs are distributions, and the mean is one of their sufficient statistics? Similarly, I did not understand, and I did not immediately see what 'phasic changes' are - what do they look like to participants - and why or how this 'kappa' 'captures sensitivity' to them. The enumerated list in this paragraph is its only saving grace. I think it is really poor practice to expect readers to refer to other publications to understand this, and this problem is particularly acute in journals that imposed space limits to such accounts.

The second point where I believe that the article fails is its emphasis on paranoia being 'domain general' rather than 'social specific'. First, there appears to be a conceptual confusion here. The sentence 'Before the pandemic, people who were more paranoid (scoring in the clinical range on standard
scales6, 8) were more likely to switch their choices between options, even following positive feedback' suggests that the results of this study are driven by people who exhibit high levels of paranoia. It is a priori unlikely that high (clinical-like) levels of any mental health symptoms will not be associated with domain-general cognitive biases, whether there is a core process which is specific to the condition. For example, if we were to take the methamphetamine animal model in the study of the authors that they cite in support, a variety of insults and psychiatric processes could set up vicious cycles disrupting dopaminergic processes, increasing paranoia. As an example, experience of high levels of sustained stress or social defeat, or usage of cannabis, etc. etc., could bring about processes wherein paranoia is the specific social-cognitive part of interacting feedback loops of cognition, behaviour and biology stabilizing both specific and non-specific processes.

It appears that the authors would like to make the point that paranoia is the downstream product of generic cognitive deficits, but do not have the evidence to support this and therefore use a language that neither pins its colours to the mast nor does it admit that the question of social vs. general cognition is too poorly specified and researched here to make a serious contribution here.

Returing to the task, there is no evidence here that the present task is particularly sensitive to social vs. generic processess. First, people are specifically told to imagine a scenario in the social task which is not particularly relevant to them. When they carry out the task, they do not believe that they are interacting with avatars. Indeed, they do not interact with the avatars at all - they just choose one or another. On the face of it, the social version of the task appears just a toy, with no ecological validity. There are indeed socially framed tasks that participants know are computer-simulated, yet they recruit socially specific effects, so the above argument does not hold universally. However, here we have no validation data to believe this is a good task for distinguishing specifically social cognitive features. The very least that would be required is that (a) the two versions of the task distinguish social vs. non-social cognition in neurotypical participants (b) that the differences are well captured by the computational model, which offers variables of good construct validity to describe the differnces with respect to processes of relevance to paranoia and (c) despite b, high paranoia does not alter the social-nonsocial difference.

There are further critiquest that should be offered to the 'generic vs social' claims of this article. A key issue is the confusion in a few parts of the article between what happens within the high-paranoia group, vs. what happens across levels of paranoia. A key example is the paragraph: 'High paranoia participants win-switched more than low paranoia participants before the lockdown (MDEMM=0.116, SEEMM= 0.031, pEMM=0.0002) and during reopening... High and low paranoia did not differ in their win-switching during lockdown (MDEMM<0.001, SEEMM= 0.027, p EMM=0.987). Again, consistent with a domain-general account, there were no differences between behaviour in the social and non-social tasks. In sum, reopening increased irrational win-switching in more paranoid participants.' This account strongly suggests that win-switching (and presumably the associated model parameters) are not an inherent characteristic of paranoia, as it did not differ during lockdown - it was 'mollified'. The text continues the confusion: '...It appears that lockdown had a mollifying effect in high paranoia, perhaps by enforcing avoidance behaviours12, decreasing social interaction and thus assuaging concerns about others'. Yet, 'concerns about others' is very much what measured paranoia is about. If 'concerns about others' was a key mediating factor, one would expect it to be reflected in lower paranoia per se, not in the cognitive processes within the high paranoia group. If the authors claimed that some other process, not plausibly reflected in the GPTS measurements, was affected by lockdown 'mollified' the effects of paranoia on cognition, that would make more sense. For example, it could be that people that *still* had high paranoia were less anxious during lockdown. Or it could be that the material conditions of their life changed. Any number of hypotheses could be put forward, of course.

The section 'The win-switch data, κ, and μ30 estimates suggest that lockdown ameliorated learning disturbances in paranoid subjects. ... proactive state lockdown ... correlated negatively with sabotage belief ... These data suggest that early and decisive state interventions may have mitigated paranoia during the escalating uncertainty of lockdown.' is similarly unclear. It is not at all clear what 'paranoia mitigation' has taken place. Is it that the entire sample showed lower levels of sabotage beliefs, in line with less GPTS paranoia? This is a first-pass meaning, but it unlikely be so, given the structure of the sample. Then, does this mean that 'paranoia mitigation' means amelioration of learning disturbances in people who have as high paranoia as the pre-lockdown ones? This appears to be more in line with the data, but can hardly be called 'paranoia mitigation'. Does 'paranoia mitigation' simply mean lower levels of sabotage beliefs in high-paranoia-scale-scoring people? If this is the case, the authors should not call it 'paranoia', because paranoia is hardly 'the belief that others bear malicious intent towards us', as per the first sentence of the paper. As famously satirized in 'Catch 22', believing that enemy troops want to kill you during war is not 'paranoia'.

If I can make a couple of suggestions, first, the authors need to explain to us a lot more clearly the direction between the inferred parameters and the descriptive statistics of the data. It is not easy to see why *win*-switch behaviour rather than *lose*-switch behaviour should be particularly sensitive to higher reward expectation. This is the kind of point where the reviewer hasn't understood what is going on and s/he frustrates the authors, but I do not understand why winning should prompt switching because of a systematic factor like reward expectation, rather than a random process (e.g. that high-paranoia participant s' difficulty in adjusting volatility down feeds into the temperature of their response function, driving 'nonsense' responses). Alternatively, I note that the reversal tasks reverse very regularly very many times in both so-called social and non-social tasks. It is possible that some entraining has taken place, or some superstitious model underlying reversals drives these changes.

Second, the authors need to tell us more about valence, long before delving into the issue of social vs. nonsocial cognition. Paranoia is strongly valenced (and so is anxiety, which the authors mention). Is there any evidence that the task show valence effects that might be relevant to paranoia? Indeed, the idea that paranoid people have a higher expectation of reward could be a model-fitting artefact (and the idea that their behaviour is explained by their being 'soon frustrated' would again intuitively be related to loss-sensitive rather than win-sensitive behaviour). In other words, how does the model behave in generative mode on task measures other than those 'dialed into' its design?

Third, the demographic data seems as the authors aknowledge, we can describe the effect of the pandemic in terms of both (i) threat and (ii) uncertainty. The task and model seem extremely good in assessing (ii). I would have liked to see a lot more about whether the population parameters talked about have more to do with a perception of uncertainty, then reflected in task parameters, rather than the complex socially mediated explanations the authors offer. What evidence is there, for example, that low mask-wearing belief in 'tight' states is an indicator of increased uncertainty, that may be reflected not only in higher paranoia but (adaptively or not?) in uncertainty-sensitive task parameters?

Finally, the period examined not only saw the pandemic, state-level responses and BLM events. It is very strange that the authors do not mention anything about how the enormous controversy and polarization associated with US national / presidencial politics in general, and responses to the pandemic specifically, may have affected both paranoia and uncertainty estimates in study participants, especially depending on their political allegiance. As QAnon have clearly suggested in analogous situations, this may not be by chance, but that this article is itself part of a nefarious liberal plot (Sorry, I coldn't resist a paranoid joke, colleagues! Actually, it's a great article!).

Reviewer #2:
Remarks to the Author:
The authors present an interesting and timely manuscript investigating the complex interactions

between the COVID-19 pandemic, paranoia, belief updating, and health behaviours. The key findings suggested that COVID-19 elevated paranoia and more erratic belief updating – while these findings are not particularly novel, a strength of the paper was the timing in which data was collected (i.e., pre-COVID, during lockdown, following reopening), which allowed for a more fine-grained analysis of the specific impact societal paranoia can have on individuals' beliefs and their uptake of health behaviours (e.g., mask wearing).

Particularly interesting was the impact of mask wearing, a public health measure to combat the spread of COVID-19, but which itself elevated paranoia, particularly in areas were adherence was low. The findings may influence social policy around lockdowns and public health interventions, given how these appear to impact paranoia. The belief updating tasks were designed well (social, non-social), and the sample was adequate, allowing for state comparisons, across different time points as the social/pandemic situation changed. The analyses were also appropriate for the questions being asked, and potential confounds were accounted for within these analyses (e.g., other events that may have caused social unrest). The conclusions were justified and consistent with the results presented, and there was adequate caution regarding the generalisability of the findings to other countries/crises.

While the manuscript is commendable for these reasons, there were also some elements which the authors may wish to address in a revision. While the justification for assessing paranoia and belief updating was touched on, there was a lack of a theoretical model guiding the hypothesis or direction of the analysis; the paper seemed to be more exploratory than driven by a theoretical framework around social paranoia (e.g., theories of belief formation, particularly paranoid beliefs – see Daniel Freeman's work on this topic). Compounding this was that there was no indication that the hypotheses or analyses were preregistered. Some variables, such as conspiracy beliefs, appeared without warning towards the end of the results sections, and were not clearly defined as factors of interest, nor was a rationale/justification provided for including this variable. The Discussion did not appear to mention other studies on the influence of public health interventions on paranoia or whether these public health practices will be upheld (e.g., do public health measures, like the compulsory wearing of face masks, typically elevate paranoia?).

Reviewer #3:
Remarks to the Author:
Review for "Paranoia and belief updating during a crisis"

Summary of review

This study leverages repeated cross-sections of pre-lockdown, lockdown and post-lockdown measures of paranoia and belief updating from an online sample of US respondents. In addition to providing estimates of how lockdown measures are associated with paranoia and related measures, the study also provides a difference-in-difference analysis of the specific effect of state-level policies on mask-wearing.

I would like to commend authors for this interesting paper, which I enjoyed reading and learned a lot from. In particular, I am impressed by the authors' use of econometric techniques developed to make causal inferences from observational studies to research psychological outcomes – something that is, unfortunately, still very rare.

However, I also believe that there are few limitations of the current version of the manuscript. Among other, more minor issues, these limitations concern the number of different hypothesis tested, the difference-in-difference analysis, and a few interpretations that are not sufficiently tied to the analysis. I will elaborate on each of these concerns, as well as several additional but minor issues below. I hope my comments are helpful for revising this interesting paper for Nature Human Behaviour or some other journal.

1. Pre-registration of analysis and multiple hypothesis testing:

a) I appreciate that different disciplines have different norms surrounding i) pre-registration of analysis plans (PAP) and ii) multiple hypothesis testing. Given the many hypotheses tested in this observational study, it would be really helpful for the credibility of this paper to know which, if any, of the analysis were pre-specified, and which were not. Also, it would be good to know which of the effect estimates remain significant once the authors adjust p-values for multiple hypothesis testing.

b) Unless the “Cultural Tightness and Looseness” analysis was pre-registered (see above), it is really difficult to know what to make of it. There are many, similarly plausible, moderators that vary at the state-level – such as the political ideology of a state, the parties in power, etc. -- and the CTL variable will also be correlated with many other variables. The paper does not explain why we should focus on CTL, and which other plausible moderators the authors have tested. Given all these concerns, the speculative interpretation on lines 316ff does not seem justified. My view is that the paper is better off without this analysis and section.

2. Difference-in-difference (DID) analysis:

a) DID models with repeated cross-sections, as used in this study, need to show that the sample composition is constant over the study period, or convincingly adjust for those differences. Figure 6 (and the associated discussion) goes some way in this direction and shows the distribution of several (pre-treatment) covariates. But rather than a few graphs, I would like to see formal placebo tests (F-test) of no differences across pre-lockdown/lockdown/re-opening samples for a much wider range of variables unaffected by COVID-19 impact and lockdown policies (an expanded version of what is provided on line 271 ff.). Note that income should not be part of this test since this variable is likely affected by the COVID-19 crisis itself.

b) The key assumption underlying DID models is parallel trends (in absence of the treatment). This assumption is, of course, untestable. But the authors should test whether trends between treated and control states did not diverge in the pre-treatment period (before onset of the mask policy). Do the authors have the pre-treatment data to conduct such a test? This would be crucial to at least indirectly validate the DID assumption. Without this tests, it is very difficult to assess the credibility of the estimate.

c) The author seems to justify (line 384ff) that DID rests on the assumption that treated and control units have similar levels of pre-treatment covariates. This is false: similar pre-treatment levels are neither a necessary nor a sufficient condition for DID. The key assumption is parallel trends (see above and/or the relevant chapter in Reference 14 of the author's manuscript).

d) The mask-wearing policy is clustered at the state level. Do the authors adjust the standard errors of the DID analysis for this clustering (e.g. using the methods developed in Caermon, Gelback and Miller, 2012, Journal fo Business & Economic Statistics)? As far as I can see, that is not done, and I suspect that the p=0.018 effect they document will turn insignificant once this clustering is taken into account.

3. Win-switch rate, mu_2 and mu_3 and many other variables are all estimated quantities, but later used as dependent or independent variables for further statistical analysis. How are these models take the estimation uncertainty of these predictors into account?

4. Several times, the study makes claims and suggests implications that are not grounded in data. A few examples: The section on public health implications on line 334 has too little grounding in any of the analysis provided in this paper. This section should either be directly tied to the analysis in this paper (or additional analysis conducted as part of the R&R) or be dropped. Similarly, the interpretation provided on lines 243-245 has to be tied more closely to the preceding analysis. Another point in case is the post-hoc rationalization of the contradictory results discussed on line 346ff. The provided justification "lockdown may have offered fewer opportunities to be caught..." sounds speculative and is not grounded in data. Generally, I would like to urge the authors to avoid any inferences that are not justified by the data and analysis provided in this study. This advice seems particularly important for studies that attempt to shed light on controversial (i.e. politicized) policies such as mask-wearing considered here.

Minor comments

- The paper contributes to our understanding of paranoia during the COVID-19 pandemic. Thus, the title should specify the type of crisis considered here, i.e., "COVID-19 crisis" or "public health crisis" or similar.

- How is "pro-active" state lockdown coded? Relative to other states? Or relative to the epidemiological situation in its own state? How robust are the results to alternative codings?

- The fonts used in some of the Figures (e.g. Figure 1 c and d) is way too small. For example, I wasn't able to read what I assume are p-values from hypothesis tests embedded in these figures – and I have a pretty big screen!

- The labeling of the figures could be better. Figure 2 and 4 should label the y-axis as expected reinforcement and volatility, respectively. Figure 3 and 4 a doesn't indicate what the star in the middle of the figure indicates (statistical significance?). The star is also a really hard to distinguish from the data points. Figure 3 b should clarify that State Proactivity is the x-axis for all the subfigure plots.

- Many of the figures should go to the online appendix as they are not relevant for the main paper, including Figures 6, 7 and maybe Figure 8.

- Period missing on line 115.

- Equation 4 defines only the coefficients, but not the variables.

- Equations (unnumbered) on lines 768 and 777 indicate that the dependent variable is a predicted quantity (hat on y). Is this correct? If so, why is there an error term in these equations?

Author Rebuttal to Initial comments

April 19th 2021

Dear Dr. Horder,

Re: Paranoia and Belief Updating During the COVID-19 Crisis

My colleagues and I were extremely grateful for the considered comments that you and the three Reviewers provided on our work. We are delighted to respond with a thorough revision of our manuscript, addressing each concern with new analyses, additional data and a comprehensive re-write of our introduction, results, and discussion. We feel the manuscript is much improved because of your excellent suggestions. We hope that you agree, and that you now consider our work suitable for publication in *Nature Human Behaviour*.

The reviewers were positively disposed toward our work. They did however have substantive concerns. We briefly summarize here, and then respond in detail – point by point – below. In brief:

Statistical rigor and inferential flexibility. We now implement false discovery rate correction for multiple statistical comparisons across all of our analyses. More broadly, the Editor and Reviewers queried whether our analyses were pre-registered. They were not. This experiment arose serendipitously – we had begun exploring social and non-social belief updating in January. As the pandemic descended, we kept recruiting, in order to explore the impact of the evolving world situation on participants' belief updating. Our a priori hypothesis – based on our prior work with this task and population - was that paranoia would increase and, choice promiscuity during reversal learning would change likewise. We did not expect the largest effects would occur at reopening. All analyses presented after that were exploratory. We acknowledge them as such clearly in the revision. That being said, we now correct for multiple comparisons and conceptually replicate the associations in our follow up data. We believe we are now appropriately circumspect in the revised version of the manuscript. We hope that the Editor and Reviewers agree that, while exploratory, our data and inferences warrant publication.

Dependent variables. Reviewers 1 and 3 were concerned about our use of paranoia as both a dependent and independent variable. Furthermore, reviewer 1 wanted us to tie the effects of policy more directly to our behaviour and self-rating results. In order to address these concerns, we now report paranoia as a dependent variable and we include proactivity of pandemic response (during lockdown and reopening) as a factor in our analyses of variance. In brief, we devised a metric of lockdown proactivity that renders a state's response more vigorous if it locked down early and remained locked down. We now report significant pandemic period by proactivity interactions for paranoia, task behaviour and model parameters. These analyses allow us to infer that paranoia, behaviour and prior beliefs about volatility varied with pandemic period and did so differently depending on the local policies regarding the pandemic.

Difference in Differences: Reviewer 3 applauded our econometric approach to inferring causality. However, they raised important concerns about the validity of our analyses. We now show that the parallel trends assumption holds in our data, and that our findings survive accounting for the possibility of clustered errors. We are grateful for the opportunity to implement these important controls.

Politics: The reviewers rightly pointed out that 2020 was not only a pandemic year, but also a year of political tumult. We included the Cultural Tightness analysis for this reason. Cultural Tightness is related to political beliefs. States that tend toward Republicanism are tighter. We now make this more explicit in the manuscript. Furthermore, we include new data from 160 participants, gathered in September 2020, assaying QAnon beliefs. QAnon is the bizarre right-wing conspiracy theory about the nefarious deep-state, powerful liberals, and the Hollywood elite abducting children to harvest adrenochrome from their brains. We find that QAnon belief correlates with paranoia, vaccine hesitancy, erratic task behavior, stronger volatility priors, and stronger priors on reward (Fig. 7 in the revised manuscript). We also confirm that QAnon belief is associated with more conservative political views and that these views similarly relate to task behavior and priors (included in the supplement).

Theoretical stance: We focus more on our theoretical motivation in the revised manuscript. It is our contention that general mechanisms in the non-social domain relate to paranoid ideation. We anchor that contention in prior work, like that of Daniel Freeman, who has shown that hasty domain general belief formation (jumping to conclusions in the beads or urns task) is related to paranoia[1]. Interactions with unfriendly virtual reality avatars induce more pronounced paranoia in people with more severe domain-general belief-updating biases[2].

We build upon that work by delineating mechanisms of domain-general belief updating and their relationship to paranoia. We now include new data demonstrating that participants believed that the non-social card decks were ministrating against them. We believe this places our work in the context of that of authors like Sarah Jane Blakemore who have shown an enhanced intentionality bias for non-social polygon stimuli[3].

Of course, the reviewers are correct – our social task may not have been social enough. We clearly acknowledge that possibility in the revision and we are much more circumspect in our conclusions about social theories. We were too strident previously. Furthermore, we address the reviewer's concerns about our modelling and the Win-Switch findings. We offer much more context in the revised manuscript. In previous work we demonstrated that our Hierarchical Bayesian model accounted best for the behavior of more and less paranoid participants. It better accounted for our data than a model that preferentially weighted positive or negative prediction errors.

Taken together, we believe, having responded to the Reviewer's constructive critiques, the manuscript is much improved. We hope that you all agree. The added data, analyses, and procedures bolster our original claims. Furthermore, we have removed points that reached beyond our data.

We conclude that the COVID-19 pandemic increased paranoia in the USA, particularly in July 2020 as people began to emerge back into their lives. We found that this change was associated with changes in task behavior – an increase in promiscuous belief updating, captured in participants' beliefs about the task and their learning rates. In exploratory analyses, we center these changes on the imposition of mask mandates, as well as on perceptions that those mandates were not being followed, particularly in states where rules are typically followed. We replicate and extend the associations between task beliefs and real-world conspiracy beliefs – incorporating data relating belief in the QAnon conspiracy to similar behaviors on the task and prior beliefs about the task.

We believe that we have addressed all of the Editor and Reviewer concerns with new data and analyses as well as extensive edits. If we have not described our responses already, they are also detailed below.

We suggest that this revised manuscript reaches the standards expected by the readership of *Nature Human Behavior* and will be of great interest to them. We hope that you agree.

I trust everything is in order, however, should you require any further information, please do not hesitate to contact me.

Kind regards and very best wishes,

Philip Corlett, Ph.D.
Associate Professor of Psychiatry and Psychology
Yale University

Detailed Responses

Editor

1) Referee #2 and #3 Preregistration (or lack of it). You should clearly state whether (and which of) your analyses were preregistered; if the work wasn't preregistered, you should clearly indicate what was

hypothesized a priori and why (refraining from hypothesizing after the fact), and which analyses were exploratory.

We now note that none of the analyses were pre-registered in the manuscript (page 2, line 119). We predicted that paranoia would increase as a function of the pandemic, and that that would be associated with changes in participant behavior and model parameters.

We did not expect that the biggest peak would arrive at reopening (rather than during lockdown) All analyses following that observation, at reopening, were exploratory – aimed at unpacking and explaining this observation. We now note this clearly in the manuscript (page 4, line 195).

2) Whether the analyses were preregistered or not, you will also need to correct for multiple comparisons as Referee #3 suggests.

All analyses are now corrected for multiple statistical comparisons using False Discovery Rate Correction. We now detail this correction approach in the manuscript (page 20, line 969).

3) Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

We believe all our text and figures are now compliant with the policies and requirements.

Reviewer #1:

1) I would like to recommend this article for publication.

This is an excellent article probing paranoia-relevant cognition in the US population before and during the COVID-19 pandemic. The article identifies specific cognitive features revealed by the computational model, which differ between high and low paranoia groups examined. These cognitive features, such as the expected reward rate that paranoid individuals have (higher than controls) manifest themselves in both the versions of the reversal task that the authors administered. In addition, the authors examined a small number of public health measures that varied across the US states whence the participants came from, especially whether states mandated or recommended the use of masks, what the levels of compliance to this measure were, the propensity of states to tightly follow rules, and how these variables related to paranoia.

Thank you. We are delighted the reviewer recommends that our work be published.

2) I think the demographic data is interesting, the analyses are adequate, and various recommendations that could be made in theory (such as the usage of longitudinal data, establishing the state vs. trait 'paranoia' involved etc., comprehensively assessing mental health symptoms and establishing rigorously the validity of the claim 'significantly. Anxiety increased ... but, the change was less pronounced ... than paranoia (Figure 2a), suggesting a particular impact of the pandemic on beliefs about others.') would be out of place in the context of a pandemic. However, I would like to point out that using unemployment as an indicator of socioeconomic distress is way too crude a measure to examine the impact of socioeconomic factors here. There are numerous much better measures of socioeconomic effects.

We agree completely that unemployment is not the only marker of socioeconomic status. We chose to use it because it was perhaps the most salient and proximal socioeconomic impact of the descending pandemic and lockdown. Many people lost their jobs. Those losses varied by state and they lessened somewhat at reopening in some states but not others.

We examined unemployment because it might have tracked the changes in paranoia that we observed. Unemployment has previously been associated with internet searches related to conspiracies. We reasoned that if people lost their job due to the pandemic restrictions and the associated economic changes, that they might blame that misfortune on some outgroup who also bore responsibility for the pandemic, or more conspiratorially, for fabricating the pandemic. We did not find evidence in favor of this assertion.
We emphasize the reasons for our choice in the revised manuscript (page 13, line 612). We also confirmed that Gini coefficients (a metric of income inequality) were not different between mask-mandate and mask-recommend states.

3) The first is in the portrayal of the hierarchical gaussian filter generative model. The paragraph 'Our generative model, the hierarchical ...' is very poor for readers who are not familiar with the HGF. It is way too brief an introduction, the reader having no picture of what the HGF really is, and why it is relevant to paranoia. Terms like 'tonic uncertainty' are not defined. It is sloppy with the terminology - how can a mean be a 'prior belief' - surely prior beliefs are distributions, and the mean is one of their sufficient statistics? Similarly, I did not understand, and I did not immediately see what 'phasic changes' are - what do they look like to participants - and why or how this 'kappa' 'captures sensitivity' to them. The enumerated list in this paragraph is its only saving grace. I think it is really poor practice to expect readers to refer to other publications to understand this, and this problem is particularly acute in journals that imposed space limits to such accounts.

We are grateful for the opportunity to describe our model more clearly (this text is now included in the revised manuscript – page 3, line 150 -, we have also included a figure of the model – panel 1c):

Probabilistic reversal learning involves decision making under uncertainty. The reasons for decisions may not be manifest in simple counts of choices or errors. By modeling participants' choices, we can estimate latent processes[4]. We suppose they continuously update a probabilistic representation of the task (a generative model) which guides their behavior[5,6]. To estimate their generative models, we identify: (1) a set of prior assumptions about how events are caused by the environment (the perceptual model) , and (2) the behavioral consequences of their posterior beliefs about options and outcomes (the response model[5,6]). Inverting the response model also entails inverting the perceptual model, and yields a mapping from task cues to the beliefs that cause participants' responses[5,6] (Figure 1c).

The perceptual model (Figure 1c) is comprised of three hierarchical layers of belief about the task, represented as probability distributions which encode belief content and uncertainty: (1) reward belief (what was the outcome?), (2) contingency beliefs (what are the current values of the options [decks/collaborators]?), and, (3) volatility beliefs (how do option values change over time?). Each layer updates the layer above it in light of evolving experiences, which engender prediction errors and drive learning proportionally to current variance. Each belief layer has an initial mean $\square^{0}$, a prior belief. □ encodes the impact of tonic or expected uncertainty on belief updating. The higher the expected uncertainty (i.e., 'I expect variable outcomes'), the less surprising an atypical outcome may be, and the less it drives belief updates ('this variation is normal'). □ captures sensitivity to perceived phasic or unexpected changes in the task. □ underwrites perceived change in the underlying statistics of the environment (i.e. 'the world is changing'), which may call for more wholesale belief revision. The layers of beliefs are summed and fed through a sigmoid response function (Figure 1c). We made the response model temperature inversely proportional to participants' volatility belief - rendering decisions more stochastic under higher perceived volatility. Using this model we have previously demonstrated identical belief updating deficits in paranoid humans and rats administered methamphetamine[7], and that this model better captures

participants' responses compared to standard reinforcement-learning models[7], including models that weight positive and negative prediction errors differently[8].

For ω_3 (tonic or expected uncertainty) we observed a main effect of group ($F_{(1, 198)}$=4.447, p=0.036, η_p^2=0.014) and block ($F_{(1, 198)}$=38.89, p < 0.001, η_p^2=0.064), but no effect of task or three-way interaction. Likewise, for μ_3^0 - the volatility prior - (group: $F_{(1, 198)}$=8.566, p=0.004, η_p^2=0.035; block: $F_{(1, 198)}$=161.845, p < 0.001, η_p^2=0.11) and κ, expected uncertainty learning rate, (group: $F_{(1, 198)}$=21.45, p < 0.001, η_p^2=0.08; block: $F_{(1, 198)}$=30.281, p < 0.001, η_p^2=0.031). We found a group effect ($F_{(1, 198)}$=12.986, p < 0.001, η_p^2=0.053) but no task, block or interaction effects on ω_2 – tonic uncertainty driven learning about rewards. Thus, we observed an impact of paranoia on behavior and model parameters that did not differ by the social or non-social framing of the task.

3) Emphasis on paranoia being 'domain general' rather than 'social specific'. First, there appears to be a conceptual confusion here. The sentence 'Before the pandemic, people who were more paranoid (scoring in the clinical range on standard scales6, 8) were more likely to switch their choices between options, even following positive feedback' suggests that the results of this study are driven by people who exhibit high levels of paranoia.

It appears that the authors would like to make the point that paranoia is the downstream product of generic cognitive deficits, but do not have the evidence to support this and therefore use a language that neither pins its colours to the mast nor does it admit that the question of social vs. general cognition is too poorly specified and researched here to make a serious contribution here.

Returning to the task, there is no evidence here that the present task is particularly sensitive to social vs. generic processes. First, people are specifically told to imagine a scenario in the social task which is not particularly relevant to them. When they carry out the task, they do not believe that they are interacting with avatars. Indeed, they do not interact with the avatars at all - they just choose one or another. On the face of it, the social version of the task appears just a toy, with no ecological validity. There are indeed socially framed tasks that participants know are computer-simulated, yet they recruit socially specific effects, so the above argument does not hold universally. However, here we have no validation data to believe this is a good task for distinguishing specifically social cognitive features.

The very least that would be required is that (a) the two versions of the task distinguish social vs. non-social cognition in neurotypical participants (b) that the differences are well captured by the computational model, which offers variables of good construct validity to describe the differences with respect to processes of relevance to paranoia and (c) despite b, high paranoia does not alter the social-nonsocial difference.

The reviewer makes an excellent point. Ultimately, our study cannot make definitive claims about the domain-general or domain-specific nature of paranoia. We have softened those claims in the revised manuscript (page 13, line 572). We acknowledge the reviewer's observation that our social task may not have been social enough in the revised discussion (page 13, lines 587-589).

We note though that there is a long history of imposing social narratives onto non-social tasks, for example Cramer et al (2002)[9]. We see our work in that vein. Furthermore, we would like to take the opportunity to share our motivation with the reviewer, since there were some concerns about our theoretical stance (or lack thereof).

Based on our prior work[7], we believe that domain-general (i.e. not specifically social) processes underpin paranoia. This is in contradistinction to models (often couched evolutionarily) that posit social specific mechanisms of paranoia. For example, enhanced evolutionary threat detection[7], particularly around coaltional cognition[10].

Our motivation for the experiment we launched in January 2020 was to render a social themed version of our probabilistic inference task. The versions were matched for difficulty, included identical incentives, and their underlying contingencies were indistinguishable.

Our hypothesis was that there would be no difference in the association with paranoia between the two tasks (social and non-social). However, the alternative hypothesis – which would favor a domain-specific account – was that paranoia would be more powerfully associated with the social task than the non-social.
We are influenced by the elegant work of Cecilia Heyes, who argues that much of what we call social cognition across species is actually driven by domain-general precision-weighted inference mechanisms[11]. Put simply, we learn about other people as if they were cues with a mean expected value, and a reliability [12]. Evidence for this type of view is extensive. Some of the most compelling comes from developmental work in humans. Human infants' domain-general associative learning abilities portend their social cognition and behaviour later in life[13]. Taken together with our own work, we feel that much of social cognition involves ill-posed and recursive inference problems. These are hard problems. They tax the inference machinery extensively. Any insults to that inference machinery will impair social inference (as well as inferences more broadly) [7].

To be clear, neither we nor Prof. Heyes disavow the presence or importance of domain-specific social mechanisms, or indeed, that there are human-specific, and extremely impactful processes of social exchange (like language, in service of communicating meta-cognitive precision for interlocution and ideally shared belief updating[12]). These are social-cognition proper. We contend that domain-specific

theorists of paranoia need to show that paranoia is particularly related to these specific mechanisms (like theory of mind).
This would have been a better experiment for us to have conducted. However, our data are still germane to the debate. The social and non-social versions were not significantly different in their relationship to paranoia, and that lack of difference was sustained across pandemic periods. These are important observations. In the revised manuscript we retain them. We describe them. We conclude that they challenge a particular domain-specific account. But we also acknowledge that approach and conclusions are not definitive, and that, in future experiments we will administer domain-specific social tasks as comparators.

We also connect our observations to those using tasks that have both social and non-social components. They typically observe a relationship between weighting of the social information and its volatility and paranoia. Although often there are no differences in these tasks and weightings between patients with schizophrenia and controls[14,15].

We acknowledge too that the Reviewer raises the work of Daniel Freeman, of which we are aware and which we admire. We characterize this work as similar to ours, since it often employs domain-general tasks – like the beads or urns problem – and relates them to paranoia [1]. Of course, Prof. Freeman makes manipulations that attempt to alter social cognition – for example through the virtual reality medium[1] – however, the end points are often more domain-general – e.g. reasoning about non-social information that requires flexible belief updating.

We now cite that work and its relationship to ours in the revised manuscript, and we thank the reviewer for making this important point (page 13, line 583).

Since our account of paranoia suggests that domain-general mechanisms give rise to domain-specific belief (about other people) it is incumbent upon us to demonstrate that association. After Freeman and others – who relate paranoia to non-social inference capacities – we too related paranoia to belief-updating during non-social probabilistic reversal learning – in our prior work and presently. We now bolster that assertion with new data. We ran a follow-up replication and extension experiment in September 2020. In this experiment, as before, following the social version, we asked participants to rate the degree that they felt sabotaged by the avatars. However, this time, we also asked the non-social participants to rate the degree to which they felt the inanimate card decks were sabotaging them. Just as in the social case, these sabotage beliefs were related to paranoia, more paranoid participants had stronger beliefs that the decks were sabotaging them (page 8, line 405).

We understand this result in the context of demonstrations of stronger agenticity experiences amongst people with higher paranoia – they imbue non-social entities with causal agency and intentions[3] (page 12, line 582)

Taken together, we agree with the reviewer that we cannot draw definitive conclusions about the domain-general versus specific nature of paranoia. It could be that all cognition is filtered through the lens of social mechanisms – as others have suggested. Parametrically manipulating the degree of social engagement and reciprocity in the tasks we present to people with paranoia will be key to delineating the precise contributions of social and non-social processes.

We humbly suggest though that non-social mechanisms have a substantial role.

We now draw this more circumspect conclusion in the revised manuscript.

4) A key issue is the confusion in a few parts of the article between what happens within the high-paranoia group, vs. what happens across levels of paranoia. A key example is the paragraph: 'High paranoia participants win-switched more than low paranoia
participants before the lockdown (MDEMM=0.116, SEEMM= 0.031, pEMM=0.0002) and during reopening... High and low paranoia did not differ in their win-switching during lockdown (MDEMM<0.001, SEEMM= 0.027, p EMM=0.987). Again, consistent with a
domain-general account, there were no differences between behaviour in the social and non-social tasks. In sum, reopening increased irrational win-switching in more paranoid participants.'

This account strongly suggests that win-switching (and presumably the associated model parameters) are not an inherent characteristic of paranoia, as it did not differ during lockdown - it was 'mollified'. The text continues the confusion: '...It appears that lockdown had a mollifying effect in high paranoia, perhaps by enforcing avoidance behaviours12, decreasing social interaction and thus assuaging concerns about others'. Yet, 'concerns about others' is very much what measured paranoia is about. If 'concerns about others' was a key mediating factor, one would expect it to be reflected in lower paranoia per se, not in the cognitive processes within the high paranoia group. If the authors claimed that some other process, not plausibly reflected in the GPTS measurements, was affected by lockdown 'mollified' the effects of paranoia on cognition, that would make more sense. For example, it could be that people that *still* had high paranoia were less anxious during lockdown. Or it could be that the material conditions of their life changed. Any number of hypotheses could be put forward, of course.

We have substantially revised the figure, results presentation, and discussion of the lockdown data (as well as reopening too).

At the reviewers' urging, we now present paranoia, task behavior, and model parameters as dependent variables in analyses of variance – examining the effect of pandemic period, policies, and their interaction (revised Figure 2).

With these new analyses, we feel more confident in asserting that more proactive lockdown (closing early and extensively) mollified paranoia and thence changed behavior and volatility priors.

Furthermore, we show (in Supplementary Figure 3) that people in more proactive lockdown and mask madate at reopening differentially impacted contamination concern.

We think this clarifies our interpretation in just the manner the reviewer requests. We hope that they agree.

5) The section 'The win-switch data, κ, and μ30 estimates suggest that lockdown ameliorated learning disturbances in paranoid subjects. ... proactive state lockdown ... correlated negatively with sabotage belief ... These data suggest that early and decisive state interventions may have mitigated paranoia during the escalating uncertainty of lockdown.' is similarly unclear. It is not at all clear what 'paranoia mitigation' has taken place. Is it that the entire sample showed lower levels of sabotage beliefs, in line with less GPTS paranoia? This is a first-pass meaning, but it unlikely be so, given the structure of the sample. Then, does this mean that 'paranoia mitigation' means amelioration of learning disturbances in people who have as high paranoia as the pre-lockdown ones? This appears to be more in line with the data, but can hardly be called 'paranoia mitigation'. Does 'paranoia mitigation' simply mean lower levels of sabotage beliefs in high-paranoia-scale-scoring people? If this is the case, the authors should not call it 'paranoia', because paranoia is hardly 'the belief that others bear malicious intent towards us', as per the first sentence of the paper. As famously satirized in 'Catch 22', believing that enemy troops want to kill you during war is not 'paranoia'. If I can make a couple of suggestions, first, the authors need to explain to us a lot more clearly the direction between the inferred parameters and the descriptive statistics of the data. It is not easy to see why *win*-switch behaviour rather than *lose*-switch behaviour should be particularly sensitive to higher reward expectation. This is the kind of point where the reviewer hasn't understood what is going on and s/he frustrates the authors, but I do not understand why winning should prompt switching because of a systematic factor like reward expectation, rather than a random process (e.g. that high-paranoia participant s' difficulty in adjusting volatility down feeds into the temperature of their response function, driving 'nonsense' responses). Alternatively, I note that the reversal tasks reverse very regularly very many times in both so-called social and non-social tasks. It is possible that some entraining has taken place, or some superstitious model underlying reversals drives these changes.

Second, the authors need to tell us more about valence, long before delving into the issue of social vs. nonsocial cognition. Paranoia is strongly valenced (and so is anxiety, which the authors mention). Is there any evidence that the task show valence effects that might be relevant to paranoia?
Indeed, the idea that paranoid people have a higher expectation of reward could be a model-fitting artefact (and the idea that their behaviour is explained by their being 'soon frustrated' would again intuitively be related to loss-sensitive rather than win-sensitive behaviour). In other words, how does the model behave in generative mode on task measures other than those 'dialed into' its design?

We are grateful for the reviewer’s careful and thoughtful engagement with our work.

Furthermore, we appreciate the opportunity to clarify.

This is not our first exploration of paranoia with this task and computational model. In our prior work we established that the hierarchical Gaussian Filter model best accounted for the behavioural differences between high and low paranoia participants[7]. In that work we compared the model to one that captures the ideas that the Reviewer posits: namely that there may be differences in sensitivity to positive and negative prediction errors in people with high paranoia. Others have explored biased belief updating (heeding positive events, downplaying negative, for example) by weighting positive prediction errors differently from negative[8]. We fit such a model to our data and found that the weightings of positive and negative errors did not explain the differences between high and low paranoia that we observed and that we replicate presently.
Next, the reviewer asks whether the model has biases baked into it, and by extension, how the Win-switch behavior emerges. The model has no such biases hard coded.
Our prior work suggests that the increase in win-switching in paranoia is associated with an increase in random responding. Following work on predator-prey interactions we measured the stochasticity of participant behavior and found it to be higher in high paranoia. When we simulated choices from the model – using the parameters estimated from high and low paranoia participants – the simulated behavior from high paranoia parameters was also more stochastic. Furthermore, these analyses, along sweeps of model parameter values, and hierarchical clustering suggest that win-switching and paranoia effects are not explicable by any one model parameter, but rather a set of common parameter differences that explained similar behavior evinced in the laboratory, online, and in a preclinical rodent model with methamphetamine exposure.

We now allude to this work more extensively the present manuscript (page 4, line 175).

Regarding our use of ‘mollify’, we now present the impact of policies at lockdown and reopening on paranoia, behaviour, and model parameters. We think that – taken together with the DiD analysis – we can cautiously claim that these policies differentially impacted beliefs and behaviours. We hope that the reviewer agrees.

5) The demographic data seems as the authors acknowledge, we can describe the effect of the pandemic in terms of both (i) threat and (ii) uncertainty. The task and model seem extremely good in assessing (ii). I would have liked to see a lot more about whether the population parameters talked about have more to do with a perception of uncertainty, then reflected in task parameters, rather than the complex socially mediated explanations the authors offer.
What evidence is there, for example, that low mask-wearing belief in 'tight' states is an indicator of increased uncertainty, that may be reflected not only in higher paranoia but (adaptively or not?) in uncertainty-sensitive task parameters?

In response to this reviewer and the other reviewers, we have radically revised the way we present our data.
In our analyses of the effects of the pandemic, we now treat paranoia as a dependent variable. We examine task behaviour and model parameters similarly.

We present the effects of pandemic period, and of policy proactivity (vigorous lockdown, mask mandates) and their interaction on paranoia, task behaviour and priors.

After appropriate correction for multiple statistical comparisons, in Figure 2b and c we now show that paranoia, win-switch rate and volatility priors were mollified by vigorous lockdown and exacerbated by mask mandates at reopening. In Figure 3b we show that task derived sabotage beliefs, win switch rates, and volatility priors differ with the vigor of a state responses. Furthermore, in Figure 4b we show that in mask mandate states, where paranoia was higher at reopening, win-switch behavior and volatility priors were also higher.

We believe these data tell a much clearer story and that we have demonstrated an impact of policies on behaviour and priors.

6) Finally, the period examined not only saw the pandemic, state-level responses and BLM events. It is very strange that the authors do not mention anything about how the enormous controversy and polarization associated with US national / presidential politics in general, and responses to the pandemic specifically, may have affected both paranoia and uncertainty estimates in study participants, especially depending on their political allegiance. As QAnon have clearly suggested in analogous situations, this may not be by chance, but that this article is itself part of a nefarious liberal plot (Sorry, I coldn't resist a paranoid joke, colleagues! Actually, it's a great article!).

We enjoyed the joke!
In addition to the pandemic, and interacting with it, 2020 was a politically turbulent year. The pandemic became a politicized issue as the November election approached.

We chose to address this in our analyses using Cultural Tight and Looseness (CTL).

We should have made this more explicit in our manuscript. We now do so in the revision.
CTL reflects the priority a given state or country places on rule following. These cultural preferences are highly (though not perfectly) correlated with political attitudes. Culturally loose states tend to vote Democrat. Culturally tight states tend to vote Republican. We now emphasize this point in our revised manuscript (page 7, line 366).

To further address how politics might have contributed to our results, we now incorporate data gathered in the latter half of the year. We assessed participant's performance on the probabilistic *reversal learning task, and we also asked them to rate their belief in the QAnon conspiracy theory. QAnon is a right-wing conspiracy theory, concerned with the ministrations of the deep-state, prominent left-wing politicians, and Hollywood entertainers. Its adherents believe that those individuals and organizations are engaged in child trafficking and murder, for the purposes of extracting and consuming the adrenochrome from the children's brains. They believe Donald Trump is part of a plan with the army to arrest and indict politicians and entertainers. We found that people who identify as Republican had stronger belief in QAnon. QAnon and paranoia more broadly are highly correlated. Furthermore, QAnon belief correlated with COVID conspiracy theorizing. Finally, QAnon endorsement correlated with the same task behaviors and parameters as paranoia (Mu3 for example, new Figure 8). Taken together, our analyses suggest that personal politics, local policies, and local political climate all contributed to paranoia and aberrant belief updating. We now make this statement more concretely in the paper and we include the data described above (page 10, line 481).*

Reviewer #2:
1) The authors present an interesting and timely manuscript investigating the complex interactions between the COVID-19 pandemic, paranoia, belief updating, and health behaviours. The key findings suggested that COVID-19 elevated paranoia and more erratic belief updating – while these findings are not particularly novel, a strength of the paper was the timing in which data was collected (i.e., pre-COVID, during lockdown, following reopening), which allowed for a more fine-grained analysis of the specific impact societal paranoia can have on individuals' beliefs and their uptake of health behaviours (e.g., mask wearing).

Particularly interesting was the impact of mask wearing, a public health measure to combat the spread of COVID-19, but which itself elevated paranoia, particularly in areas were adherence was low. The findings may influence social policy around lockdowns and public health interventions, given how these appear to impact paranoia. The belief updating tasks were designed well (social, non-social), and the sample was adequate, allowing for state comparisons, across different time points as the social/pandemic situation changed. The analyses were also appropriate for the questions being asked, and potential confounds were accounted for within these analyses (e.g., other events that may have caused social unrest). The conclusions were justified and consistent with the results presented, and there was adequate caution regarding the generalisability of the findings to other countries/crises.

We are extremely grateful for the positive assessment and clear, succinct, summary of our work.

2) While the justification for assessing paranoia and belief updating was touched on, there was a lack of a theoretical model guiding the hypothesis or direction of the analysis; the paper seemed to be more exploratory than driven by a theoretical framework around social paranoia (e.g., theories of belief formation, particularly paranoid beliefs – see Daniel Freeman's work on this topic).
Compounding this was that there was no indication that the hypotheses or analyses were preregistered. Some variables, such as conspiracy beliefs, appeared without warning towards the end of the results

sections, and were not clearly defined as factors of interest, nor was a rationale/justification provided for including this variable.

This is fair. We have now thoroughly revised the manuscript to make our position clearer (e.g. page 12, lines 572-589). We also elaborate further below. However, briefly, we believe that domain-general belief updating mechanisms can explain a substantial amount of paranoia. This is in-line with Professor Freeman's work, which demonstrates hasty non-social belief updating (for example during the Beads or Urns task) relates both to paranoia and to the effects of his virtual reality manipulations.

We hope that our position is much clearer in the revised version of the manuscript.

Furthermore, we now explicitly acknowledge the exploratory nature of our analyses – unpacking the impact of mask mandates on paranoia for example.

3) The Discussion did not appear to mention other studies on the influence of public health interventions on paranoia or whether these public health practices will be upheld (e.g., do public health measures, like the compulsory wearing of face masks, typically elevate paranoia?).

The data that relate real-world uncertainty to paranoia and conspiracy theorizing are somewhat anecdotal and largely historical. For example, the conspiratorial anti-semitic belief that Jewish people were poisoning wells and causing the Black Death[16]. The AIDS epidemic was associated with a number of conspiracies related to public health measures – but less directly. For example, people believed (and some continue to believe) that HIV was created (either intentionally or accidentally) through the polio vaccination program in Africa[17]. More broadly, the early phases of the epidemic were associated with heightened paranoia concerning homosexuals and intravenous drug users[18].

These examples are different from our observations though.

Perhaps the closest relative to our mask mandate result involves seatbelt laws[19]. Like masks in a viral pandemic, seatbelts are (and continue to be) extremely effective at preventing serious injury and death in road traffic accidents[20]. However, the introduction of State Laws prescribing that they are worn was associated with public outcry[19]. People were concerned about the imposition on their freedom[19]. They complained that seatbelts were particularly dangerous when cars accidentally entered bodies of water. People feared that drivers wearing seatbelts would drown. The evidence shows, first that such accidents are extremely rare, and second, that when they do happen, seatbelt wearing is not associated with excess fatality. Yet people protested for years. We do not know that their self-rated paranoia increased, but we can reasonably infer that it did (these points are now made in the discussion, page 11, line 511).

Finally, early in the pandemic Daniel Freeman and colleagues reported that pandemic conspiracy theorizing was associated with higher paranoia and less adherence to public health countermeasures[21]. We now cite this work in our revised manuscript. Our work replicates and extends these preliminary *observations from the UK by suggesting underlying mechanisms (ranging from the individual to broader cultural influences) and making cautious causal claims with regards to the impact of mandates on paranoia. We now cite this work in our revision (page 10, line 461).*

Reviewer #3:
This study leverages repeated cross-sections of pre-lockdown, lockdown and post-lockdown measures of paranoia and belief updating from an online sample of US respondents. In addition to providing estimates of how lockdown measures are associated with paranoia and related measures, the study also provides a difference-in-difference analysis of the specific effect of state-level policies on mask-wearing.

I would like to commend authors for this interesting paper, which I enjoyed reading and learned a lot from. In particular, I am impressed by the authors' use of econometric techniques developed to make causal inferences from observational studies to research psychological outcomes – something that is, unfortunately, still very rare.

We are thankful for the reviewer's positive assessment of our work

1. Pre-registration of analysis and multiple hypothesis testing:
a) I appreciate that different disciplines have different norms surrounding i) pre-registration of analysis plans (PAP) and ii) multiple hypothesis testing. Given the many hypotheses tested in this observational study, it would be really helpful for the credibility of this paper to know which, if any, of the analysis were pre-specified, and which were not. Also, it would be good to know which of the effect estimates remain significant once the authors adjust p-values for multiple hypothesis testing.

Our experiments and analyses were not pre-registered. We appreciate that this is best practice. However, in our case we did not do it, as we believed it was not possible to pre-register after we had begun acquiring the data. We began our online experiment in January 2020 prior to the declaration of the pandemic by the WHO. We had no foreknowledge of the impact the pandemic would have (indeed, in March 2020, we had the impression that we would shelter in place for just a few short weeks). We were curious what impact the developing pandemic would have on social and non-social belief updating and so we continued to gather data online. We have since learned that analyses can be pre-registered after data acquisition, but before analysis. We now acknowledge that we did not do this in the discussion. In future we will endeavor to pre-register our predictions. That being said, these data build directly on our prior publication with the probabilistic reversal task and paranoia[7]. We predicted that the pandemic would increase paranoia, and in so doing, it would change participant's behavior on the task, in a manner that was related to their paranoia. Furthermore, we correct our analyses for multiple statistical comparisons where appropriate using False Discovery Rate correction[22].

b) Unless the "Cultural Tightness and Looseness" analysis was pre-registered (see above), it is really difficult to know what to make of it. There are many, similarly plausible, moderators that vary at the state-level – such as the political ideology of a state, the parties in power, etc. -- and the CTL variable will also be correlated with many other variables. The paper does not explain why we should focus on CTL, and which other plausible moderators the authors have tested. Given all these concerns, the speculative interpretation on lines 316ff does not seem justified. My view is that the paper is better off without this analysis and section.

We acknowledge the reviewer's concerns here. However, we prefer to retain the analyses. We think that they are informative with regards to unpacking why a mask mandate might have increased paranoia. We now clearly mark the analysis as ***exploratory*** *in the revised manuscript. Cultural tightness reflects the extent to which a state's citizens tend to value rule following. The analysis connects our belief updating and paranoia data to the violation of social norms. We think this is a worthy explanation of our findings, but – at the reviewer's behest – we emphasize the exploratory nature of this analysis.*

2. Difference-in-difference (DID) analysis:
a) DID models with repeated cross-sections, as used in this study, need to show that the sample composition is constant over the study period, or convincingly adjust for those differences. Figure 6 (and the associated discussion) goes some way in this direction and shows the distribution of several (pre-treatment) covariates. But rather than a few graphs, I would like to see formal placebo tests (F-test) of no differences across pre-lockdown/lockdown/re-opening samples for a much wider range of variables unaffected by COVID-19 impact and lockdown policies (an expanded version of what is provided on line 271 ff.). Note that income should not be part of this test since this variable is likely affected by the COVID-19 crisis itself.

Thank you. We understand the concern. We thought that we had addressed it by including the CloudResearch Data – which speak to the sample from which we could have drawn. We now confirm with F-tests that neither gender nor age, nor race, nor education level, nor employment, nor income, nor medication, nor mental and neurological health differed by sampling period in the samples that we ascertained either (page 9, line 434). We hope that these additional analyses assuage the reviewer's concerns.

b) The key assumption underlying DID models is parallel trends (in absence of the treatment). This assumption is, of course, untestable. But the authors should test whether trends between treated and control states did not diverge in the pre-treatment period (before onset of the mask policy). Do the authors have the pre-treatment data to conduct such a test? This would be crucial to at least indirectly validate the DID assumption. Without these tests, it is very difficult to assess the credibility of the estimate. The author seems to justify (line 384ff) that DID rests on the assumption that treated and control units have similar levels of pre-treatment covariates. This is false: similar pre-treatment levels are neither a necessary nor a sufficient condition for DID. The key assumption is parallel trends (see above and/or the relevant chapter in Reference 14 of the author's manuscript).

The reviewer is absolutely correct. A DiD is only valid if the treatment groups share parallel trends prior to the intervention. Many DiD analysts confirm the parallel trend assumption by inspection. We wanted to assign a statistic to our decision. In the revised manuscript we confirm the parallel trends assumption by computing λ *(now described in Supplementary Figure 5). We assume that parallel trends hold.*

One issue with analyses that assert parallel trend assumptions is that they are not robust to considerations of baseline demographic differences between the treatment groups[23]. Thus, we additionally retain all of the analyses we conducted to demonstrate that cases, deaths, and unemployment, amongst other features were not different pre-intervention, nor did they change post-intervention. These analyses increase our confidence in the DiD and in our conclusion that mask mandates appear to have increased paranoia.

d) The mask-wearing policy is clustered at the state level. Do the authors adjust the standard errors of the DID analysis for this clustering (e.g. using the methods developed in Caermon, Gelback and Miller, 2012, Journal fo Business & Economic Statistics)? As far as I can see, that is not done, and I suspect that the p=0.018 effect they document will turn insignificant once this clustering is taken into account.

As the reviewer suggests, the residuals within-state may be correlated. That is, paranoia levels might be similar in people from the same state for reasons other than the policy intervention. Reviewer 3 suggested we use cluster-robust estimation. However, when the number of groups is small (less than 50), variance estimates tend to be biased towards zero, resulting in confidence intervals that are too tight. Instead we implemented a non-parametric cluster bootstrap procedure in R, which is theoretically robust to heteroscedasticity and arbitrary patterns of error correlation within clusters, and to variation in error processes across clusters. Following Cameron & Miller (2015)[24], the procedure reassigns entire states to either treatment or control and recalculates the treatment effect in each reassigned sample, generating a randomization distribution. The cluster-adjusted standard error equals 0.217. When we use that standard error to compute a p-value, we find p=0.038. Thus, with cluster corrected standard error, the effect of mask mandate on paranoia remained statistically significant.

3. Win-switch rate, mu_2 and mu_3 and many other variables are all estimated quantities, but later used as dependent or independent variables for further statistical analysis. How are these models take the estimation uncertainty of these predictors into account?

The Hierarchical Gaussian Filter approach yields estimates of the precision of parameter values for each participant. In order to address the Reviewer's important point, we incorporated these estimates into every analysis in which we explore the relationship between a parameter value and some other feature (like lockdown rigor – for example). This entailed two types of analysis: First, where ever we previously fit a regression model with an HGF parameter, we now report inverse-variance weighted regression analyses. Note that in the figures depicting these analyses, the sizes of the data-points convey the precision of the estimate for that participant. When we make comparisons of the parameter values (say between mask mandate states and mask recommend states) we compute inverse-variance weighted t-

tests. The relationships and differences we reported previously became stronger as a result of this addition. We of course apply false discovery rate correction to these analyses. Again, our findings survive this correction.

4. Several times, the study makes claims and suggests implications that are not grounded in data. A few examples: The section on public health implications on line 334 has too little grounding in any of the analysis provided in this paper. This section should either be directly tied to the analysis in this paper (or additional analysis conducted as part of the R&R) or be dropped. Similarly, the interpretation provided on lines 243-245 has to be tied more closely to the preceding analysis. Another point in case is the post-hoc rationalization of the contradictory results discussed on line 346ff. The provided justification "lockdown may have offered fewer opportunities to be caught…" sounds speculative and is not grounded in data. Generally, I would like to urge the authors to avoid any inferences that are not justified by the data and analysis provided in this study. This advice seems particularly important for studies that attempt to shed light on controversial (i.e. politicized) policies such as mask-wearing considered here.

We have removed this post-hoc speculation from the discussion

Minor comments

- The paper contributes to our understanding of paranoia during the COVID-19 pandemic. Thus, the title should specify the type of crisis considered here, i.e., "COVID-19 crisis" or "public health crisis" or similar.

We have made this change. Thank you!

How is "pro-active" state lockdown coded? Relative to other states? Or relative to the epidemiological situation in its own state? How robust are the results to alternative codings?

State proactivity is now a much larger part of our analysis. We revisited our coding of proactivity and made sure that early lockdown was weighted similarly to later reopening and that both were contributing to our calculation (see Equation 1). In response to the reviewers' query, we tested the relationship between our participants' responses and early lockdown alone as well as longer lockdown alone. Neither of these simpler features related to our participants' behaviour.

- The fonts used in some of the Figures (e.g. Figure 1 c and d) is way too small. For example, I wasn't able to read what I assume are p-values from hypothesis tests embedded in these figures – and I have a pretty big screen!

We recreated all of the figures and increased the font size.

- The labeling of the figures could be better. Figure 2 and 4 should label the y-axis as expected reinforcement and volatility, respectively. Figure 3 and 4 a doesn't indicate what the star in the middle of the figure indicates (statistical significance?). The star is also a really hard to distinguish from the data points. Figure 3 b should clarify that State Proactivity is the x-axis for all the subfigure plots.

We have remade all the figures and rewritten all figure legends. We hope they are clearer now.

- Many of the figures should go to the online appendix as they are not relevant for the main paper, including Figures 6, 7 and maybe Figure 8.

Figures 6 and 7 are now part of the supplement.

- Period missing on line 115.

This has been remedied. Thank you.

- Equation 4 defines only the coefficients, but not the variables.

Thank you – we now define the variables.

- Equations (unnumbered) on lines 768 and 777 indicate that the dependent variable is a predicted quantity (hat on y). Is this correct? If so, why is there an error term in these equations?

This was an error – we have removed the error term.

1 Freeman, D., Pugh, K., Vorontsova, N., Antley, A. & Slater, M. Testing the continuum of delusional beliefs: an experimental study using virtual reality. *J Abnorm Psychol* **119**, 83-92, doi:10.1037/a0017514 (2010).

2 Pot-Kolder, R., Veling, W., Counotte, J. & van der Gaag, M. Self-reported Cognitive Biases Moderate the Associations Between Social Stress and Paranoid Ideation in a Virtual Reality Experimental Study. *Schizophr Bull* **44**, 749-756, doi:10.1093/schbul/sbx119 (2018).

3 Blakemore, S. J., Sarfati, Y., Bazin, N. & Decety, J. The detection of intentional contingencies in simple animations in patients with delusions of persecution. *Psychol Med* **33**, 1433-1441, doi:10.1017/s0033291703008341 (2003).

4 Corlett, P. R., Fletcher, P.C. Computational Psychiatry: A Rosetta Stone linking the brain to mental illness. *Lancet Psychiatry* (2014).

5 Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience* **5**, 39, doi:10.3389/fnhum.2011.00039 (2011).

6 Mathys, C. D. *et al.* Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in human neuroscience* **8**, 825, doi:10.3389/fnhum.2014.00825 (2014).

7 Reed, E. J. *et al.* Paranoia as a deficit in non-social belief updating. *Elife* **9**, doi:10.7554/eLife.56345 (2020).

8 Lefebvre, G., Nioche, A., Bourgeois-Gironde, S. & Palminteri, S. Contrasting temporal difference and opportunity cost reinforcement learning in an empirical money-emergence paradigm. *Proc Natl Acad Sci U S A* **115**, E11446-E11454, doi:10.1073/pnas.1813197115 (2018).

9 Cramer, R. E. *et al.* Human agency and associative learning: Pavlovian principles govern social process in causal relationship detection. *Q J Exp Psychol B* **55**, 241-266, doi:10.1080/02724990143000289 (2002).

10 Raihani, N. J. & Bell, V. An evolutionary perspective on paranoia. *Nat Hum Behav* **3**, 114-121, doi:10.1038/s41562-018-0495-0 (2019).

11 Heyes, C. & Pearce, J. M. Not-so-social learning strategies. *Proceedings. Biological sciences / The Royal Society* **282**, doi:10.1098/rspb.2014.1709 (2015).

12 Heyes, C., Bang, D., Shea, N., Frith, C. D. & Fleming, S. M. Knowing Ourselves Together: The Cultural Origins of Metacognition. *Trends Cogn Sci* **24**, 349-362, doi:10.1016/j.tics.2020.02.007 (2020).

13 Reeb-Sutherland, B. C., Levitt, P. & Fox, N. A. The predictive nature of individual differences in early associative learning and emerging social behavior. *PLoS One* **7**, e30511, doi:10.1371/journal.pone.0030511 (2012).

14 Henco, L. *et al.* Bayesian modelling captures inter-individual differences in social belief computations in the putamen and insula. *Cortex* **131**, 221-236, doi:10.1016/j.cortex.2020.02.024 (2020).

15 Henco, L. *et al.* Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. *PLoS Comput Biol* **16**, e1008162, doi:10.1371/journal.pcbi.1008162 (2020).

16 Cohn, N. *The Pursuit of the Millenium*. (Oxford University Press, 1961).

17 Worobey, M. *et al.* Origin of AIDS: contaminated polio vaccine theory refuted. *Nature* **428**, 820, doi:10.1038/428820a (2004).

18 Gonsalves, G. & Staley, P. Panic, paranoia, and public health--the AIDS epidemic's lessons for Ebola. *N Engl J Med* **371**, 2348-2349, doi:10.1056/NEJMp1413425 (2014).

19 Giubilini, A. & Savulescu, J. Vaccination, Risks, and Freedom: The Seat Belt Analogy. *Public Health Ethics* **12**, 237-249, doi:10.1093/phe/phz014 (2019).

20 Robertson, L. Road death trend in the United States: implied effects of prevention. *J Public Health Policy* **39**, 193-202, doi:10.1057/s41271-018-0123-2 (2018).

21 Freeman, D. *et al.* Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychol Med*, 1-13, doi:10.1017/S0033291720001890 (2020).

22 Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat Med* **9**, 811-818, doi:10.1002/sim.4780090710 (1990).

23 Jaeger, D. A., Joyce, T.J., Kaestner, R. A Cautionary Tale of Evaluating Identifying Assumptions: Did Reality TV Really Cause a Decline in Teenage Childbearing? *Journal of Business & Economic Statistics* **38** (2020).

24 Cameron, A. C., Miller, D.L. A Practitioner's Guide to Cluster-Robust Inference. *The Journal of Human Resources* **50**, 317-372 (2015).

Decision Letter, first revision:

Our ref: NATHUMBEHAV-210113930A

22nd June 2021

Dear Dr. Corlett,

Thank you for submitting your revised manuscript "Paranoia and belief updating during the COVID-19 crisis" (NATHUMBEHAV-210113930A). It has now been seen by the original referees and their comments are below. I apologize for the delay in this last round of review.

As you can see, the reviewers find that the paper has improved in revision.

I am therefore happy to say that we accept in principle to publish it in Nature Human Behaviour, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and we will send you a checklist detailing our editorial and formatting requirements shortly. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,
Jamie

Dr Jamie Horder
Senior Editor

Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

Thank you for the opportunity to review this paper again.
I think it has improved still more. The discsussion of explanatory claims both with respect to cogntion and with respect to the benefits and limitations of methodology (pre-registration, Difference-in-Difference analyses, clearer account of external events such as the run-up to the US presidential election) are now more in proportion.

The theoretical discussion regarding (so-called) social-specific vs. generic cognition is now clearer too, and I think it will form the basis of much future theoretical clarification and empirical research.

I think there may be a small error in line 256 - the date 11/02/20 should probably refer to 2021.

I would be grateful for the publication of this work without delay, so that I can share it with interested juniors and colleagues.

Reviewer #2 (Remarks to the Author):

The authors have adequately addressed the comments raised by the reviewers.

Reviewer #3 (Remarks to the Author):

Second review for “Paranoia and belief updating during the COVID-19 crisis”

I appreciate all the changes that the authors made in response to the feedback received from me, the reviewers, and the editor. I agree with the authors that these revisions have substantially improved the paper.

My initial comments focused mostly on the preregistration, multiple hypothesis testing, and the difference in difference analysis. I will discuss those in turn:

Preregistration: The revised manuscript now clarifies that the study has not been pre-registered and should therefore be considered “exploratory” in nature.

Multiple hypothesis testing: some of the analysis now use the Benjamini-Hochberg procedure to adjust for multiple hypothesis testing.

Difference-in-difference analysis: The revised manuscript provides a series of tests, which increases trust that the parallel trend assumption holds in this context. The description in the text is, however, not entirely accurate (line 196). Note that the parallel trend assumption states that the change in potential outcomes under control between the pre- and post-treatment period is the same for treated and control units. Since we cannot observe the counterfactual potential outcome under control for treated units in the post-treatment period, we cannot directly test this assumption. What we can do (and the authors do) is test whether trends developed in parallel in the pre-treatment period. If we do not find any evidence of divergent trends in the pre-treatment period, this increases our trust in the parallel trend assumption in the post-treatment period. However, this is not the same as a direct test of the parallel trend assumption.

Decision letter, final requests:

** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors. **

Our ref: NATHUMBEHAV-210113930A

22nd June 2021

Dear Dr. Corlett,

Thank you for your patience as we've prepared the guidelines for final submission of your Nature Human Behaviour manuscript, "Paranoia and belief updating during the COVID-19 crisis" (NATHUMBEHAV-210113930A). Please carefully follow the step-by-step instructions provided in the attached file, and add a response in each row of the table to indicate the changes that you have made. Ensuring that each point is addressed will help to ensure that your revised manuscript can be swiftly handed over to our production team.

We would like to start working on your revised paper, with all of the requested files and forms, as soon as possible (preferably within two weeks). Please get in contact with us if you anticipate delays.

When you upload your final materials, please include a point-by-point response to any remaining reviewer comments.

If you have not done so already, please alert us to any related manuscripts from your group that are under consideration or in press at other journals, or are being written up for submission to other journals (see: https://www.nature.com/nature-research/editorial-policies/plagiarism#policy-on-duplicate-publication for details).

Nature Human Behaviour offers a Transparent Peer Review option for new original research manuscripts submitted after December 1st, 2019. As part of this initiative, we encourage our authors to support increased transparency into the peer review process by agreeing to have the reviewer comments, author rebuttal letters, and editorial decision letters published as a Supplementary item. When you submit your final files please clearly state in your cover letter whether or not you would like to participate in this initiative. Please note that failure to state your preference will result in delays in accepting your manuscript for publication.

In recognition of the time and expertise our reviewers provide to Nature Human Behaviour's editorial process, we would like to formally acknowledge their contribution to the external peer review of your manuscript entitled "Paranoia and belief updating during the COVID-19 crisis". For those reviewers who give their assent, we will be publishing their names alongside the published article.

<b>Cover suggestions</b>

As you prepare your final files we encourage you to consider whether you have any images or illustrations that may be appropriate for use on the cover of Nature Human Behaviour.

Covers should be both aesthetically appealing and scientifically relevant, and should be supplied at the best quality available. Due to the prominence of these images, we do not generally select images featuring faces, children, text, graphs, schematic drawings, or collages on our covers.

We accept TIFF, JPEG, PNG or PSD file formats (a layered PSD file would be ideal), and the image should be at least 300ppi resolution (preferably 600-1200 ppi), in CMYK colour mode.

If your image is selected, we may also use it on the journal website as a banner image, and may need to make artistic alterations to fit our journal style.

Please submit your suggestions, clearly labeled, along with your final files. We'll be in touch if more information is needed.

<b>ORCID</b>

Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Nature Human Behaviour has now transitioned to a unified Rights Collection system which will allow our Author Services team to quickly and easily collect the rights and permissions required to publish your work. Approximately 10 days after your paper is formally accepted, you will receive an email in providing you with a link to complete the grant of rights. If your paper is eligible for Open Access, our Author Services team will also be in touch regarding any additional information that may be required to arrange payment for your article. Please note that you will not receive your proofs until the publishing agreement has been received through our system.

Please note that <i>Nature Human Behaviour</i> is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. <a href="https://www.springernature.com/gp/open-research/transformative-journals"> Find out more about Transformative Journals</a>

<B>Authors may need to take specific actions to achieve <a href="https://www.springernature.com/gp/open-research/funding/policy-compliance-faqs">

compliance</a> with funder and institutional open access mandates.</b> For submissions from January 2021, if your research is supported by a funder that requires immediate open access (e.g. according to <a href="https://www.springernature.com/gp/open-research/plan-s-compliance">Plan S principles</a>) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route our standard licensing terms will need to be accepted, including our <a href="https://www.springernature.com/gp/open-research/policies/journal-policies">self-archiving policies</a>. Those standard licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

For information regarding our different publishing models please see our <a href="https://www.springernature.com/gp/open-research/transformative-journals"> Transformative Journals </a> page. If you have any questions about costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com.

Please use the following link for uploading these materials:
[REDACTED]

If you have any further questions, please feel free to contact me.

Best regards,
Chloe Knight
Editorial Assistant
Nature Human Behaviour

On behalf of

Jamie

Dr Jamie Horder
Senior Editor
Nature Human Behaviour

Reviewer #1:
Remarks to the Author:
Thank you for the opportunity to review this paper again.
I think it has improved still more. The discsussion of explanatory claims both with respect to cogntion and with respect to the benefits and limitations of methodology (pre-registration, Difference-in-Difference analyses, clearer account of external events such as the run-up to the US presidential election) are now more in proportion.

The theoretical discussion regarding (so-called) social-specific vs. generic cognition is now clearer too, and I think it will form the basis of much future theoretical clarification and empirical research.

I think there may be a small error in line 256 - the date 11/02/20 should probably refer to 2021.

I would be grateful for the publication of this work without delay, so that I can share it with interested juniors and colleagues.

Reviewer #2:
Remarks to the Author:
The authors have adequately addressed the comments raised by the reviewers.

Reviewer #3:
Remarks to the Author:
Second review for “Paranoia and belief updating during the COVID-19 crisis”

I appreciate all the changes that the authors made in response to the feedback received from me, the reviewers, and the editor. I agree with the authors that these revisions have substantially improved the paper.

My initial comments focused mostly on the preregistration, multiple hypothesis testing, and the difference in difference analysis. I will discuss those in turn:

Preregistration: The revised manuscript now clarifies that the study has not been pre-registered and should therefore be considered “exploratory” in nature.

Multiple hypothesis testing: some of the analysis now use the Benjamini-Hochberg procedure to adjust for multiple hypothesis testing.

Difference-in-difference analysis: The revised manuscript provides a series of tests, which increases trust that the parallel trend assumption holds in this context. The description in the text is, however, not entirely accurate (line 196). Note that the parallel trend assumption states that the change in potential outcomes under control between the pre- and post-treatment period is the same for treated and control units. Since we cannot observe the counterfactual potential outcome under control for treated units in the post-treatment period, we cannot directly test this assumption. What we can do (and the authors do) is test whether trends developed in parallel in the pre-treatment period. If we do not find any evidence of divergent trends in the pre-treatment period, this increases our trust in the parallel trend assumption in the post-treatment period. However, this is not the same as a direct test of the parallel trend assumption.

Final Decision Letter:

Dear Dr Corlett,

We are pleased to inform you that your Article "Paranoia and belief updating during the COVID-19 crisis", has now been accepted for publication in Nature Human Behaviour.

Before your manuscript is typeset, we will edit the text to ensure it is intelligible to our wide readership and conforms to house style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see http://www.nature.com/nathumbehav/info/gta). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

Please note that <i>Nature Human Behaviour</i> is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. <a href="https://www.springernature.com/gp/open-research/transformative-journals"> Find out more about Transformative Journals</a>

<B>Authors may need to take specific actions to achieve <a href="https://www.springernature.com/gp/open-research/funding/policy-compliance-faqs"> compliance</a> with funder and institutional open access mandates.</b> For submissions from January 2021, if your research is supported by a funder that requires immediate open access (e.g. according to <a href="https://www.springernature.com/gp/open-research/plan-s-compliance">Plan S principles</a>) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route our standard licensing terms will need to be accepted, including our <a href="https://www.springernature.com/gp/open-research/policies/journal-policies">self-archiving policies</a>. Those standard licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at <a href="https://www.nature.com/reprints/author-reprints.html">https://www.nature.com/reprints/author-reprints.html</a>. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

We look forward to publishing your paper.

With best regards,
Jamie

Dr Jamie Horder
Senior Editor
Nature Human Behaviour

P.S. Click on the following link if you would like to recommend Nature Human Behaviour to your librarian http://www.nature.com/subscriptions/recommend.html#forms

** Visit the Springer Nature Editorial and Publishing website at <a href="http://editorial-jobs.springernature.com?utm_source=ejP_NHumB_email&utm_medium=ejP_NHumB_email&utm_campaign=ejp_NHumB">www.springernature.com/editorial-and-publishing-jobs</a> for more information about our career opportunities. If you have any questions please click <a href="mailto:editorial.publishing.jobs@springernature.com">here</a>.**