

Jan-Philipp Fränken
Stanford University
Department of Psychology
450 Jane Stanford Way
Bldg 420
Stanford, CA 94305
Email: jphilipp@stanford.edu

August 4, 2023

Dear Professor Oaksford,

We appreciate the time taken by you and the two reviewers in assessing our manuscript and we are grateful for the valuable feedback. In response to your comments, we made the following major changes:

- We have restructured and revised our entire manuscript. We now include detailed descriptions of our Task (p. 6), Computational Models (p. 7), and Experiments (p. 11) within the main text.
- We have restructured and increased the size of our figures and updated all captions. Specifically, we increased the number of figures from three to six in the main text and provide revised supplementary figures in the appendix.
- We have addressed all other comments raised by yourself and the two reviewers.

We provide point-by-point responses and descriptions of our revisions below. Our responses to comments are marked in **blue**. References to updated passages in the manuscript are marked in **green**. Moreover, we have attached a “diff” file as part of our submission which shows the changes made to the original manuscript.

We feel this major revision has helped us to greatly improve the paper, and we are grateful for having been given the opportunity to do so. We hope that the revised version of the manuscript is considered fit for publication in *Cognition*.

Thank you very much for your time and consideration.

Kind regards,

J.-Philipp Fränken, Simon Valentin, Chris Lucas, and Neil Bramley

Editor’s comments: Comment 1 I have now received two reviews from experts in the area. Both reviewers saw value in the research you report in this paper. However, both reviewers felt that the short report format led to too telegraphic a report of your studies. They provide many suggestions for changes to your manuscript all of which will mean putting material from the supplementary materials into the main text.

Response 1: Thanks, yes with hindsight, we agree that readers will need more details upfront to follow the experiments and results. As such, we have restructured and expanded the main text in several places. This included moving supplementary material to the main text and expanding on it in various places. Concretely:

- We now offer detailed explanations of our Task (p. 6), Computational Models (p. 7), and Experiments (p. 11) in the main text.
- Additionally, using a new high level overview Figure 1 (p. 4), revised Figure 2 (p. 6), and new Figure 3 (p. 8), we provide a detailed illustration of our task and inference setup. Furthermore, we have included the following sections of our manuscript in the new Task section starting on p. 6:

In our social learning task, participants have to combine their own private evidence samples from the environment with other agents’ judgments to make their own judgments (see Fig. 1a & Fig. 2 for an illustration. See Fig. A-1 and our online demo for full instructions). Under our minimal cover story, participants have crash landed on one of two planets with different proportions of blue ■ and red ■ fish. Participants are told that on the first planet, Planet Blue, $\frac{2}{3}$ of the fish are blue, and $\frac{1}{3}$ of the fish are red. On the second planet, Planet Red, proportions are reversed. Aside from the different proportions of red and blue fish, the planets are indistinguishable. Over the course of ten trials participants sample private evidence by fishing, occasionally catching either a blue (■) or red (■) fish and so learning about the proportion of fish on the current planet. The probability with which participant catch a blue or red fish versus no fish (i.e., no evidence marked as ■) at each trial is unknown to participants. (p. 6)

[...]

Importantly, the private evidence collected by other agents is *never* observed by participants. Moreover, participants don’t know about the frequency with which other agents observe private evidence, nor do they know at which trial other agents observe private evidence. The only thing participants know is that other agents are on the same planet, and hence the evidence sampled by the other two agents must come from the same distribution. To succeed in the task, participants must thus infer the summative impact of the unknown private evidence seen by the other two agents from observations of the other agents’ public judgments and combine it with their own private evidence. Participants have to provide a judgment about the planet they are on repeatedly across a series of ten trials, using a seven-point scale ranging from 3–blue (highly confident planet blue) to 3–red (highly confident planet red). This in turn forms the social evidence that other participants see (depending on their position in the network). Participants can always see both their own history of private evidence samples, their own previous judgments, but crucially also

the previous judgments of neither, one, or both other agents depending on their position in the network. (p. 6)

I fully agree with this verdict and to this end, I have spent some time going through your supplementary materials. This has led me to require the following clarifications in addition to those proposed by the reviewers. I was unclear on how the experiment unfolded from the main text, and this was not made clearer in the supplementary material. I found the following sentence confusing:

“Here, Chris can see the evidence of both Neil and Simon. Simon sees Neil’s judgment but can’t see Chris’. Neil can see neither Chris’ nor Simon’s evidence”

There’s a difference between having access to the other agents’ evidence versus access to their judgments based on that evidence. Clearly, if a participant has access to the evidence the other agents collect, then they should incorporate that evidence directly into their judgments (assuming the conjugate beta distribution, by just adding to the a and b parameters). Consequently, this sentence is confusing as it oscillates between seeing evidence vs seeing judgments based on evidence. Chris seems to see all the evidence collected by everyone, but Simon only sees Neil’s judgments based on what evidence Neil himself collected. Neil cannot see the evidence on which Chris or Simon base their judgments, nor, I assume, their judgments. This sentence, therefore, seems to confound your findings as participants are likely as confused by it as I am.

Response 2: Apologies for this confusing sentence. We have removed it from the manuscript and clarified throughout the manuscript when we are referring to private evidence from the environment (draws of blue or red fish) versus social evidence in the form of other players’ likert-scale judgments. We now explicitly state that participants *never* see the private evidence seen by other participants but must infer this evidence from the observed judgments provided by other agents (see Response 1).

Regarding the potential for this wording undermining participants’ understanding of the task, we believe this was not the case because participants went through a thorough interactive instruction phase introducing them to all elements of the task before testing them on their understanding in a comprehension quiz. Crucially also, this sentence was phrased with the word “judgments” rather than “evidence” in the task avoiding this ambiguity (the use of evidence in “[...] Chris can see the evidence...” was a typo, the actual sentence shown to participants during instructions was “[...] Chris can see the judgments...”; see Figure R-1). Additionally, we employed graphical cues (green arrows) to visualize the communication pathways (i.e., exchanges of judgments between participants) and observation of private evidence. Furthermore, we supplemented these cues with a specific hint provided in the instructions visible to participants:

NOTE: You can never see other players’ catches. You can only see your own catches and other players’ judgments to revise your own judgments. Other players always provide accurate judgments, as they also get bonuses for guessing correctly in each round.

We apologise for not including a screenshot of this hint in our initial submission. We have now corrected the link to our online [demo](#) and included a new Figure A-1 (p. 26) showing the main

instruction phase screens, which includes the above hint. While we recognise that some of our instructions' wording might not have been ideal in hindsight, we included so much redundancy between visual and text cues and comprehension questions that we believe participants had a good understanding of the task. For your reference, screenshots from our online demo have been included as Figures R-1 and R-2 below.

Your job is to work out which planet you are at over the course of your fishing trip. To do this you will need to pay attention to whatever fish you catch yourself (you will get 10 attempts to fish) and you will also be able to draw on the judgments both the other players.

Task

You have 2 minutes and 32 seconds left to complete this page.

While it will be clear whose judgments you can see, **you can't be sure what the other players can see.** That is, you cannot be sure whether player 2 can see player 3's judgments or if player 3 can see player 2's judgments or if either player can see your own judgments. Below is an illustration showing one possibility. Here, Chris can see the judgments of both Neil and Simon. Simon sees Neil's judgment but can't see Chris'. Neil can see neither Chris' nor Simon's judgments.

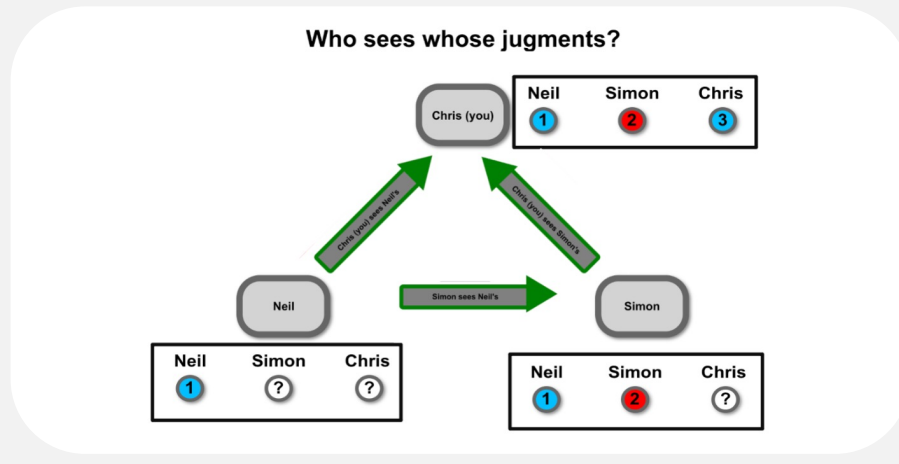


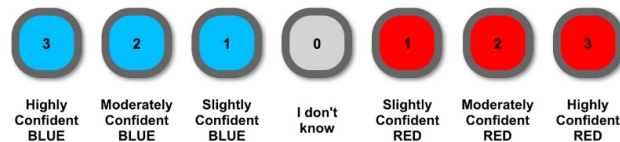
Figure R-1: Screenshots of instructions from our online demo including illustration of who can see whose judgments.

Procedure

You have 1 minutes and 9 seconds left to complete this page.

In each of the 10 'rounds' you will:

1. Check to see if you caught a fish, and what colour it is if you did.
2. Provide a **judgment about which planet you are at using the response format below.**



3. See the judgments of the other players. Like your own judgments these will range between 'Highly confident RED' and 'Highly confident BLUE' (see below). **Always consider current and past evidence** (fish / other players' judgments) when making a new judgment.

Neil



Chris



Simon



You will earn a bonus of £0.15 for selecting the correct planet in each round. To select the correct planet, you need to consider *both* other players' judgments and your own fish.

NOTE: You **can never** see other players' catches. You can only see your own catches and other players' judgments to revise your own judgments. Other players **always** provide accurate judgments, as they also get bonuses for guessing correctly in each round.

Next

Figure R-2: Screenshots of instructions from our online demo including explicit hint that participants can never see the private evidence (fish catch) from other agents (highlighted using magenta box).

Comment 2 I assume from SI Figure 1 that all participants see are the judgments made by the other agents responding to the evidence they collect. It is important that participants do not believe that the other agents' judgments are in response to the same sequence of evidence that they are seeing. If they believe it is the same sequence, then any variation from their own judgment of which planet they all are on will provide evidence of the other agents' unreliability, and consequently, they should ignore the evidence of the other agents' judgments. The trial summary in SI Figure 1 does give the impression that each agent is responding to the same sequence of evidence. Is there somewhere in the instructions where it is made clear that this is or is not the case? Moreover, if they all are collecting their own evidence independently (so each agent is seeing a different stream of evidence) and just have access to each other's judgments, disagreements would seem to have to be based on the non-uniform distribution of fish of different colours between the various places at which each agent is fishing. In this case, the judgments are not particularly informative and drilling down to aggregate the evidence on which they are based is the only sensible procedure. In summary, I am concerned that this statistical task is not relevantly similar to aggregating testimony say in legal proceedings, where collusion between witnesses (dependence) may lead one to discount their evidence. You need an argument that your task is relevantly similar to situations where independence/dependence matters. You also need to fully describe the experimental procedure and sequence of events in the main text.


Response 3: As we explain above in Response 2, participants knew the other agents could not directly observe each other's private evidence but that all agents had to integrate their own private evidence with other agents' judgments. To do that normatively, they should reverse engineer their best guess of the evidence going into other agents' judgments under consideration of the judgments' dependence, and then combine this inferred evidence with their own private evidence to make accurate judgments.

As such, we do think that our task is relevantly similar to, for example, aggregating testimony in legal settings. Specifically, participants had to aggregate a series of judgments ("testimony") from two potentially dependent agents ("witnesses"). Experiment 1 is designed to manipulate this dependence while holding the surface evidence constant. The simulated agents' judgments were the same in all three conditions, but the private evidence our participants ought to infer from this evidence according to our Level-2 model depended on what we told them about the network structure: In condition two, participants were told that agent C could see agent B's judgments meaning C's judgments in spite of B's suggest C caught lots of red fish, while in condition three the reverse inference holds, it is most likely that B has caught lots of blue fish. Consequently, the directional inferences for Level-2 inference differed substantially between conditions (see Fig. 4a and Fig. A-2a, panel *C sees B known*, cyan line).




To better bring these design choices out, we have revised our Task and Experiments sections (see Response 1). Moreover, we added the following: [...] Specifically, our task enables us to manipulate the social information network—who actually sees whose judgments—as well as manipulating participants beliefs about the parts of the network they do not observe directly—whether other agents can see their judgments, or each other's judgments. The true structure influences how information propagates and as such participants structure beliefs influence the weights they should assign to the judgments of other agents (i.e., discount or increase the inferred evidence behind

other agents' judgments). For example, if agent A believes that agent C sees agent B's judgments, they should anticipate that there is likely to be some redundancy whereby C's judgments are partly a consequence of B's (akin to our coffee shop example from the introduction). If C makes similar judgments to B perhaps they have not seen any evidence of their own. Alternatively, if C's judgments differ dramatically from B's (Fig. 2, left), this would suggest that C has observed substantial evidence, enough to override the influence of judgments from B. Furthermore, if the distal network structure is unknown to an agent, there is the potential for them to infer it by recognizing patterns of inheritance (i.e., if C's judgment reliably shifts in line with B's preceding judgment). In our experiments, we either explicitly provide the network structure to participants (Experiment 1, conditions two and three, and Experiment 2) or withhold structure information (Experiment 1, condition one, Experiment 3). In conditions where we withhold structure information, at the end of the task we have participants provide a structure judgment (Fig. 2, right). This setup enables us to study whether participants aggregate judgments at their face value, versus additionally consider and accommodate potential redundancies implied by the communication structure, and further whether they can infer this structure from the communication sequences. We next formalize these intuitions using a nested set of computational models. (p. 7)

In the Stimuli for Experiment 1 we now include:

In all three conditions, participants played the role of agent A and caught a single red fish  themselves on trial 2 (t_2) and no fish on any other trial. In all three conditions the participants also saw the same sequence of opposing judgments from B and C (Fig. 2, left), which, depending on the relationship between B and C, led to qualitatively different predictions for Level-2 inferences.¹ We selected a judgment sequence for agents B and C to produce a large qualitative difference for Level-2 predictions between conditions. When generating stimuli, we assumed that agents B and C were reliable (c.f. Hawthorne-Madell & Goodman, 2019), meaning that their provided judgments corresponded to those with a high posterior probability under their respective beliefs h . The resulting judgment sequences for B and C are shown in Fig. 2, left. We note that there is no ground truth in this condition as we picked B's and C's judgments in advance to separate both model predictions within a condition as well as a given model's predictions between conditions. (p. 12)

In describing the stimuli for Experiment 2, we now say:

Similarly to Experiment 1, the participant playing agent A observed one red fish  at the second trial. Meanwhile, the participant playing agent B observed one blue fish  on trial three, and agent C observed one red fish  on trial seven. The staggered and alternating private evidence seen by players A, B and C was selected to produce differences in the judgments of agent A depending on both the structure condition and whether they performed predominantly naïve Level-1 or Level-2 inferences. To unpack why this is the case, recall that agent B does not get any social evidence, and agent C gets only evidence about B's earlier judgments in condition two (C sees B known)

¹In the experiment, the roles of red and blue were randomized between subjects and the simulated agents' judgments were reversed accordingly.

meaning only agent A needs to worry about dependence between judgments. Given this setup, we anticipate that B will make blue-leaning judgments in both conditions since all they see is one blue fish. In condition two, Agent C will regard B's blue-leaning judgments as initial evidence favoring the blue planet before their own red catch leaves them finally with a roughly neutral judgment. If A overlooks this dependency (naïve inference), they miss the chance to deduce that C's latterly neutral judgment is actually most compatible with them having seen a red fish. As a result, a naïve agent A will consider their own red fish and B's blue judgment as providing insufficient evidence for either planet, leading to A holding a neutral final judgment in condition two. In the independent structure condition one C is likely to make red-leaning judgments and so A should also favor red on the balance of evidence. However, if A correctly identifies the structure and adjusts for the dependency between B and C in a rational, Level-2 manner, we would expect A to infer the additional red fish from C's neutral judgments in condition two, resulting in preference for the red planet in *both* conditions. (p. 15)

Comment 3 *Your figures (which require 300% magnification to read on a computer screen) plus captions are not sufficient for Cognition.*

Response 4: Thank you for pointing this out; we agree that the figures were too compact. In response, we have updated and greatly enlarged our figures, and also revised the captions, including descriptions of x-axes and y-axes while keeping descriptions of the actual task in the main text (with the exception of the new Fig. 1 which is meant to provide a high-level summary of our paper). Our new figures are as follows:

- Our new Figure 1 (p. 4) provides a high level overview of our task, models, and the three experiments.
- In response to Reviewer 1, our revised Figure 2 (p. 6) moved from the SI file to the main text.
- Our new Figure 3 (p. 8) presents a more comprehensive illustration of our beta-binomial model, showing how beliefs are updated using private evidence, as well as how beliefs map to discrete judgments probabilities on our scale, and then how judgments, such as “1-red” are used to infer an agent’s underlying belief, and thus estimate the unobserved private evidence it is based on.
- We now present the results for each experiment in separate figures:
 - Our new Figure 4 (p. 12) shows the results from Experiment 1, which now includes further details on participant structure judgments, as requested by Reviewer 1.
 - Our new Figure 5 (p. 16) shows the results from Experiment 2.
 - Our new Figure 6 (p. 17) presents the results from Experiment 3, providing additional details on structure judgments as requested by Reviewer 1.
- Beyond revising all figures in the main text, we have updated our key supplementary figures and included these in the appendix of our manuscript (Fig. A-1 on p. 26, Fig. A-2 on p. 27,

and Fig. A-3 on p. 28). The only remaining content in the SI file are the supplementary tables. We have removed raw judgments from Experiment 1 from our SI as these were always the same for simulated agents B and C and are now shown in Fig. 2 (left).

Comment 4 In summary, you need to address these issues and all the issues raised by the reviewers in your paper or cover letter. I am willing to give you the space to resubmit as a full article to facilitate this as I believe your results could be theoretically significant. Upon receiving your revision, I will send it out to the same two reviewers, and, hopefully, decide on receiving their reviews.

Response 5: Thank you again for this opportunity. We hope that the revised version of our manuscript matches your and the reviewers' expectations.

Reviewer 1's comments:

Comment 1 The authors explore how people gather information by combining their own sampled evidence with social evidence—both what their peers decide and what network structure might have led to their peers' decisions. The authors use an impressive breadth of experimental methods, from very controlled experiments that hold constant computer "peers" behavior at each trial, to triadic experiments that bring three live participants together within multiple network structure. Furthermore, each experiment is paired with systematic modeling that complement the study's narrative. Overall, I think this is neat work that I would love to see published. My main concern is that it packs a lot into a brief report. I believe the inclusion of results that are currently in the supplementary material and some additional discussion would greatly improve the overall quality of this paper.

Response 6: Thank you for your positive and constructive feedback. We have now restructured our manuscript and provide detailed descriptions of our task, models, and experiments in the main text (see Response 1). We have also updated all figures (see Response 4).

Comment 2 The paper could be improved by laying out more context for the results and conclusions. Specifically, using individual level analyses, the authors find converging quantitative evidence that people (sub-optimally) ignore network information in the $B \rightarrow C$ and $C \rightarrow B$ conditions. Instead most participants behaved as though B and C independently made decisions. Potential concerns for this broad claim is that it largely depends on (1) the extent to which participants can correctly interpret the task instructions explicitly stating what the network structure is, and (2) whether participants believe their peers are actually being influenced by the other. Maybe participants look like Level-1 thinkers because they were not convinced their peers' decisions were conditioned on the other peer. In fact, in experiment 1, B and C both behave as though they may be downright ignoring the other as they adamantly stick to saying they are in the same "world" across trials. Moreover, the main text currently excludes experiment 2 and 3 trial by trial data, which makes it hard for readers to judge how natural the computer behavior in experiment 1 is. Under what circumstances in the triadic experiments do participants, in their actual behavior, downright ignore their (single) peers' input (with players $B \rightarrow C$) and so extending that social assumption to others is justified (A to assume C is ignoring B)? Either inclusion of more results from the supplementary material or a discussion point can help resolve this potential concern about the conclusions.

Response 7: Thank you for raising these issues. Several concerns were shared by the editor so Responses 1–3 explain why we are confident participants understood the task. We now provide more detail about the instructions Fig. A-1 (p. 26) and include a new task explanation section (p. 6). Additionally, we have moved our visualizations of the raw trial-by-trial data from Experiments 2–3 into the main manuscript (Fig A-3, p. 28). Moreover, we extended the discussion of our results with a new paragraph on differences between experiments (p. 18):

An important consideration in interpreting the above findings is the different nature of the social evidence (i.e., judgments by B and C) between Experiments 2–3 and Experiment 1. The social evidence in Experiment 1 was simulated to align with rational and reliable inference patterns (see

Equation A-2), while in Experiments 2–3, the judgments by B and C were made by actual human participants. Human judgments varied across triads, often diverging from the patterns predicted by rational simulations (see Fig. A-3). While this discrepancy had no direct influence on our ability to characterize agent A's (participant) inferences as predominantly naïve, it significantly affected the performance of the structure-sensitive Level-2 learner. Specifically, Level-2 predictions diverged directionally from those of an omniscient observer of all the caught fish (see Fig. 5a and Fig. 6a). Notably, in Experiment 3 condition two, where the social network structure was undisclosed and agent C could see agent B's judgments, Level-2 inferences in the role of agent A led to predictions that were directionally wrong on average, slightly favoring the blue planet (Fig. 6a) despite the overall evidence observed by all agents being two red fish ■ versus one blue fish ■. This was due to Level-2 inference incorrectly assuming that the human agents were also behaving like rational utility maximizing learners, while model fitting suggests this was only true for a small proportion of participants. Diverging from Level-2's predictions, participants playing agent A made judgments actually aligned more closely with the ground truth, favoring the red planet in both conditions of Experiments 3. This is consistent with the idea that participants relied on simpler accumulation strategies, leading to inferences more robust to the complexity and ambiguity inherent in social evidence.

Regarding your concern (2) whether participants believe their peers are actually being influenced by the other. Well this is really at the heart of the social inference phenomena we are investigating: We are interested in to what degree there is a match or mismatch between what people actually take away from peers' judgments, and what that take-away implies about what their peer's take away from from one another's judgments. The two problems are inherently recursively interrelated. It is clear, for example, that Agent A is actually influenced by Agents B and C, this is why we see departures from level 0 inference across all experiments when modelling Agent A's judgments. Past work, including some of our own, has shown that people can sometimes infer communication structure from sequences of judgments, and are capable of adjusting their inferences somewhat based on this at least when reasoning from simulated agents or in single shot settings (Fränken, Valentin, Lucas, & Bramley, 2021; Whalen, Griffiths, & Buchsbaum, 2018). Indeed, the existence of this sensitivity is supported by our main effect in Experiment 1. However the idea that everyone reverse engineers and adjusts for these dependencies in real social settings where the dependencies can quickly become unmanageably deep and complex seems like a far stronger claim. It was not at all clear to us that level-2 + social inferences are really the dominant mode for repeated social information exchanges. Indeed, our Experiments 2 and 3 show that, when people are forced to make judgments repeatedly in this low bandwidth setting, they predominantly fall back more on simple aggregation of each others judgments which curiously and somewhat paradoxically means people implicitly assume others judgments are uninfluenced by one another even while being influenced themselves. We comment on this curious pattern in the discussion.

Comment 3 The inferences that participants make about the social network structure is intriguing, and they validate how participants interpreted the network structure. Thus, I would love to see that in some form in the main text!

Response 8: Thank you for this suggestion! We have included two new figures (Figure 4, p. 12

and Figure 6, p. 17) in the main text, showing participants' average structure judgments. See Response 3 as well as the following addition in the Stimuli section of Experiment 1:

When generating stimuli, we assumed that agents B and C were reliable (c.f. Hawthorne-Madell & Goodman, 2019), meaning that their provided judgments corresponded to those with a high posterior probability under their respective beliefs h . The resulting judgment sequences for B and C are shown in Fig. 2, left. We note that there is no ground truth in this condition as we picked B's and C's judgments in advance to separate both model predictions within a condition as well as a given model's predictions between conditions. (p. 12)

and the following new section to the results sections for Experiment 1 and Experiment 3:

In condition one with no structure information, we wanted to ensure that the simulated judgment sequences for B and C made it difficult to infer the existence or direction of any dependency between B and C since this would undermine the instruction manipulation. To test the degree to which the sequence of simulated judgments were informative about the communication structure, we present participants' structure judgments in condition one alongside posterior probabilities for each connection under a Bayesian structure learning model; see Structure Learning for details; Fig. 4c). This reveals that a normative structure learner assigns a low probability to a dependent communication structure being behind this sequence, assigning a probability of 0.03 to C sees B and 0.21 to B sees C. Participants' edge selections were also lower than random chance at 22% and 37%, suggesting that they could not tell whether or not there was a dependency between B and C. Overall, the above edge selections primarily functioned as a sanity check and were deemed sufficiently low for our manipulation to be effective. Interestingly, this analysis did reveal that both participants and our structure learner frequently and erroneously hallucinate that either or both of B and C were reacting to their judgments (i.e., those of agent A). (Experiment 1, p. 13)

and

Finally, examining the quality of the structure judgments, the proportion of participants' marking each edge is shown in Fig. 6c and compared against the marginal posterior edge probabilities under our Bayesian structure learning model. As expected, the proportion of participants marking an edge from B to C was higher in condition two (60% for participants and 75% for the Bayesian structure learner) than in condition one in which B and C were independent (31% for participants and 60% for a normative structure learner). Notably though, the accuracy of the structure judgments by both participants and the normative structure learner were low. Both participants and the model often wrongfully judged that their own (agent A's) judgments were visible to agents B and C. (Experiment 3, p. 18)

Comment 4 Additionally, I feel that perhaps more background is needed to justify the inclusion of "sticky" models. And, some clarification about the model: are "sticky" judgments assuming that participants themselves are sticky? Or does level 2 sticky judgments mean that participants are also assuming their peers are sticky?

Response 9: Thank you for raising this point. Our "Sticky" judgments models simply assume

that the judgments of the participant being modelled tend to be sticky, but do not additionally assume stickiness in the judgments of the other participants. Our computational models section (starting on p. 7) now includes a new paragraph called “Accommodating Autocorrelation: ‘Sticky’ Models” (p. 10) in which we discuss our motivation for including sticky models:

In our task, participants are required to make the same judgment ten times as evidence arrives (Fig. 2, left). This setup presents a challenge for the above inference models, which predict judgments at a specific time point are based on the total evidence (both observed and inferred), and are independent of the participant’s previous judgments. A wealth of research shows that people’s responses when probed repeatedly tend to be autocorrelated over and above what is licensed by the evidence. This has been shown in both single learner (e.g., Bramley, Dayan, Griffiths, & Lagnado, 2017; Dasgupta, Schulz, Tenenbaum, & Gershman, 2020; Hogarth & Einhorn, 1992; Lieder, Griffiths, Huys, & Goodman, 2018) and multi-agent settings (e.g., Fränken, Theodoropoulos, & Bramley, 2022). To partially accommodate such order effects, and thereby enhance our models’ ability to capture the meaningful patterns in participants’ judgment sequences, we thus incorporated an additional variant (“family”) of the inference models. In addition to using the cumulative density $I_x(\alpha, \beta)$ (Equation 1) to assign discrete probabilities to each judgment $y_i \in \{y_1, y_2, \dots, y_k\}$ followed by a softmax function, the second family of models incorporates an additional free parameter (mixture weight π), which mixes each soft-maxed model prediction with the possibility of simply “sticking” to the previous judgment (see Additional Model Details). Lastly, we include a random baseline model that predicts each judgment with a uniform probability of $\frac{1}{k}$ ($k = 7$), resulting in a total of seven competing models for each experiment.²

Comment 5 I was confused about how to interpret the “True State” in Figure 3.

Response 10: We have updated our figure captions in our new figures 4–6 and removed “True State” from Figure 4 since as you note, there was no real ground truth in Experiment 1. For the other two figures, we have changed the label to “Combined Private Evidence” which is hopefully more intuitive. “Combined private evidence” refers the predictions of an omniscient learner that had direct access to all private evidence observed by agents.

Comment 6 Perhaps some reorganization would help readers understand the task and results. For example, the authors could move the concrete predictions for behavior (Expt 3) before the individual modeling section (Expt 2).

Response 11: Thank you for this suggestion! We hope that the new structure of our manuscript will better help readers understand our task and results (see Response 1 and Response 3).

Re: *For example, the authors could move the concrete predictions for behavior (Expt 3) before the individual modeling section (Expt 2):* For each experiment, the concrete predictions for behavior are now described before discussing the results from our individual modeling analysis and Figures 4–6 have been restructured such that the behavioral results and predictions appear above the modeling results.

² “Sticky” models were added as exploratory analysis after our data collection was complete. Note also that our “sticky” models do not additionally assume stickiness in inverting the evidence behind other agents’ judgments, see Discussion.

Comment 7 The link to the demo experiment is broken.

Response 12: Thank you for pointing this out. Our link should work now. Here is it again for your reference: [link](#).

Reviewer 2's comments:

Comment 1 I appreciate the argument and topic of the paper. The experimental design is reasonable, as the revised urn scenario provides a controlled setting for testing how people receive information and update their beliefs. As you can see below, I believe there is a lot of positive aspects of the manuscript, but I would encourage the authors to set out the motivation for the hypotheses more clearly, expand the results from experiment 2, and provide more direct model justifications. I believe this can be done and therefore recommend that the manuscript is revised and resubmitted.

Response 13: Thank you for your encouraging feedback! We have thoroughly revised and restructured our manuscript and figures to better motivate our task, models, and experiments (see Response 1, Response 3, and Response 10). We provide more detailed responses to your specific comments below.

Comment 2 I would encourage the authors to be more specific in how they describe the motivation for the experimental hypotheses. In particular, the sticky model provides an anchor, which is a reasonable assumption. However, it is not clear why this particular model is suggested compared with a model that includes, for example, confirmation biases. This is not to say that the authors should necessarily run competing model comparisons with other naïve strategies, but rather to motivate the selection of this naïve strategy.

Response 14: We have revised our motivation for our models in several paragraphs, as addressed in Response 1, Response 9 and Response 18. Regarding your concern: *It is not clear why this particular model is suggested compared with a model that includes, for example, confirmation biases*: Several recent papers (e.g., Fränken, Theodoropoulos, Moore, & Bramley, 2020; Hawthorne-Madell & Goodman, 2019; Whalen et al., 2018) have studied social reasoning within a rational-speech-act framework (Frank & Goodman, 2012) where one can articulate inferences of different recursion depths ("levels"). Building on this tradition and inference framework (see paragraph 2 on p.5: *To test this and understand how participants make sense of the social evidence, we will analyze behavior using three nested computational models (Fig. 1b), which, akin rational speech act models (c.f. Frank & Goodman, 2012), assume different levels of recursion: [...]*), we were interested in understanding to what degree previous rational/deep recursive inferences such as Whalen et al. (2018) hold true in more realistic settings requiring multiple judgments with simulated (Experiment 1) and real (Experiments 2–3) social peers. Moreover, confirmation bias is usually more relevant in larger discrete hypothesis spaces where evidence pertaining to one hypothesis might be uninformative for distinguishing between several others. Here, we only have two hypotheses (Planet Blue versus Planet Red), so if the belief in one planet goes up, the other goes down. We acknowledge that the present setting might involve several other heuristics such as copying judgments or simply averaging judgments which have not been fully explored, despite these being potential additional reasons for making people more "naïve".

Comment 3 In general, for the experimental design, the hypotheses and theoretical motivations could be presented more clearly. In particular, it would be beneficial to present competing predictions from the models as well as the theoretical rationale that underpins these differences. At

present, the experimental design is not explained fully. This is more a matter of clarification than of change, however.

Response 15: Thank you for pointing this out, both the editor and first reviewer made a similar point. As such, we now provide an detailed description of our task (p. 6), computational models (p. 7), and experimental procedures (p. 11) in the main text and have revised our figures to better illustrate our setup. See Response 1 and Response 3 for quoted text and details of these changes.

Comment 4 It is not clear whether the study was pre-registered. If not, it would be good to clarify the potential exploratory nature of the analyses. For example, experiment 1 suggests that L1 sticky explanations are more appropriate for the largest part of the sampled population. However, it is not clear whether this was predicted a priori to conducting the experiment or if this is merely observed a posteriori.

Response 16: Our study was not pre-registered. We now state this on p. 11: The experiments were not preregistered. Based on our previous work (Fränken et al., 2020, 2021), we hypothesized that participants' judgments setting would be best described by a naïve, Level-1 account. and a footnote on p. 11: "Sticky" models were added as exploratory analysis after our data collection was complete. Note also that our "sticky" models do not additionally assume stickiness in inverting the evidence behind other agents' judgments, see Discussion.

As such our analyses are partly exploratory but also partly motivated by a-priori predictions (Fränken et al., 2021). Specifically, our a-priori prediction was that participants' judgment patterns would fall short of full (here level-2) accommodation of dependencies in the evidence. The present experiments were effectively a detailed follow-up of our initial CogSci paper (Fränken et al., 2020) and poster (Fränken et al., 2021), and we predicted that Level-1 to win across experiments as in Fränken et al. (2021). Given our recent work on order effects in sequential judgments (Bramley et al., 2017; Fränken et al., 2022) as well as strong evidence for order effects more generally (Hogarth & Einhorn, 1992; Lieder et al., 2018), we expected that sticky variants of our models would be better at characterising participants' judgment sequences.

Comment 5 The results of experiment 1 suggests that 'the majority of participants neglecting the dependency between sources'. It would be good to discuss the effect of this - would the same intuition hold for social contexts where dependency is a clear epistemic guide? For example, a person (A) may recommend a restaurant and add that another person (B) also likes the place. If the recipient then meets B who reports about the restaurant, would the naïve strategy also predict that the recipient should neglect the dependency between the sources?

Response 17: Indeed, this is the naïve prediction. The idea behind the naïve model is that any sort of communication signal, irrespective of the underlying evidence (i.e., whether private/first-hand evidence or hearsay), is integrated at its face value. So hearing about the same source twice (once indirectly through A and once directly from B) would result in counting the evidence twice. Of course, every model has its limits and here we are focused on low bandwidth signals where there is little opportunity for a signaller to flag that part of their judgment is based on another. Natural language communication, as you point out, opens the door to making more

explicit delineations that we would expect. Moreover, while it seems plausible that people would adjust for these more explicitly flagged evidential redundancies, there is also a lot of evidence across psychology that people struggle to adjust completely even when they know they should (e.g., Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Zajonc, 1968)

To address your comment, we have updated our General Discussion on p. 19 as follows: [...] A possible implication of this is that individuals may not only discount or inflate the evidence from dependent sources when directly hearing from them, but also when receiving information about dependent sources' opinions indirectly. For instance, if person B recommends a restaurant and mentions that person C also enjoys it, a naïve recipient A, upon meeting B and hearing about the restaurant, may mistakenly update their beliefs based on B's report.

Comment 6 The results from experiment 2 could be expanded, as it is not clear whether the same analyses from experiment 1 would replicate in 2 regarding the between-subjects manipulation. Further, it would be useful to reflect on the differences in outcome, as L1 fits go from 48% to 62%. This comment extends to Experiment 3, where L1 fits 41% of the responses. **Response 18:** We have now expanded the results sections for both Experiment 2 (p. 15) and Experiment 3 (p. 16). Additionally, we have incorporated a new section discussing differences across experiments (refer to p. 18 and Response 7).

Regarding the variation in outcomes (48%, 62%, and 41% best fit by the sticky Level-1 model across experiments), we argue that this discrepancy can be ascribed to the fact that participants knew the network structure in Experiment 2 *and* at the same time faced noisy real-world agents whose judgments might have been inconsistent with the known structure (as compared to Experiment 3 where the structure was unknown or Experiment 1 where the other two agents were simulated and provided more consistent judgment patterns). This inconsistency may have caused participants to trust the other agents' judgments from Experiment 2 less and thus they may have behaved more "sticky", resulting in 62% best predicted by Level-1 sticky.

Comment 7 Given the proximity of the results to Hahn et al (2020) that you cite in the references, I would encourage a direct comparison with their results. If memory serves me right, they find that the structure of the network impacts the evidence that a person can sample. Your results suggest that assuming peers are rational can lead to systematic judgment errors - Hahn et al (2020) find that the structure of the network influences rational agents' capacity to know the true distribution of evidence. Are there any substantive differences in cognitive assumptions that would account for the differences between your findings and those of Hahn and colleagues? Do you see your results as complementary or in opposition to their findings?

Response 19: Thank you for raising this point. We agree that Hahn et al.'s results are highly relevant to this study and have now updated our discussion on p. 18 to include a direct comparison as follows: [...] A key feature of our analyses and results is that the structure of a social network has the potential to shape the beliefs of members, often away from the ground truth. As such, we view our results as complementary to prior simulation-based studies that demonstrate this distorting effect of social network structure in simulations (Fränken & Pilditch, 2021; Hahn, Hansen, & Olsson, 2020; Lewandowsky, Pilditch, Madsen, Oreskes, & Risbey, 2019; Madsen,

Bailey, & Pilditch, 2018; Madsen & Pilditch, 2018). These studies simulate how information propagates through large artificial networks, typically assuming that social communications are received and integrated accurately but “naïvely” (in our terminology), albeit exploring how they may be tempered by agent-specific considerations like trust. For example, Hahn et al. (2020) show that both dense connectivity and clustering in artificial social networks reduces the “truth tracking” of information propagation among otherwise rational agents. Here we show that even in minimal three-person social networks with known structure, the kinds of naïve inference that produce these distortions is the dominant behavior of human social reasoners. [...] In addition to citing Hahn et al. (2020), we included Madsen et al. (2018) and Fränken and Pilditch (2021) which are additional simulation-based studies showing how the structure of a social network in combination with simple peer-to-peer transmission of information can result in beliefs that can diverge arbitrarily far from the true evidence.

Comment 8 In the introduction, you mention social learning failure leads to poor outcomes - it is not clear if the argument is that these outcomes are due to people lacking social learning (which I presume is not the case with the spread of misinformation), if it is due to poor quality of cues in those instances, or if it is some other failure that you mean. It would be useful to spell out in more detail what the causal chains are for these cases (p. 2, paragraph 1).

Response 20: We agree that this was not clear enough in the previous version of the manuscript and have revised paragraph 2 on p. 3 to accommodate your feedback: When it falls short, social learning can lead to poor collective outcomes. For instance, repeated sharing of redundant information between agents can result in echo chambers and misinformation cascades (Fränken & Pilditch, 2021; Jasny, Waggle, & Fisher, 2015). Likewise, persistent exposure to harmful or inaccurate content can precipitate moral outrage (Brady, McLoughlin, Doan, & Crockett, 2021; Crockett, 2017), endorse science denial (Scheufele, Hoffman, Neeley, & Reid, 2021), amplify political polarisation (Levin, Milner, & Perrings, 2021; Tokita, Guess, & Tarnita, 2021), and trigger financial bubbles and crashes (De Martino, O’Doherty, Ray, Bossaerts, & Camerer, 2013).

Comment 9 I appreciate the limitations brought forth in the paper, as these are relevant to the design of the experiment and the inferences drawn from the data. The authors may also consider including perceived reliability of sources as a possible extension, as this has been shown to influence belief revision and dependency judgments - see e.g. the following:

Bovens, L. & Hartmann, S. (2003). Bayesian epistemology, Oxford, UK: Oxford University Press.

Hahn U., Harris A. & Corner A. (2009). Argument content and argument source: An exploration, *Informal Logic* 29, 337-367.

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2015). The Appeal to Expert Opinion: Quantitative support for a Bayesian Network Approach, *Cognitive Science* 40, 1496-1533.

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2020). The impact of partial source dependence on belief and reliability revision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46 (9), 1795-1805.

Response 21: Thank you for suggesting these references and the idea including perceived source reliability, we have added these to the following revised sentence on p. 19: Another limitation of the present work is that we did not account for participants' perceived credibility or reliability estimates of other players, which has previously been shown to influence social belief revisions and dependency judgments (Bovens, Hartmann, et al., 2003; Hahn, Harris, & Corner, 2009; Harris, Hahn, Madsen, & Hsu, 2016; Madsen, Hahn, & Pilditch, 2020). Additional modeling extensions could thus probe or manipulate learners' beliefs about the expertise or trustworthiness of peers' judgments to better understand how people weigh private versus social evidence.

References

- Bovens, L., Hartmann, S., et al. (2003). *Bayesian epistemology*. Oxford University Press on Demand.
- Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33), eabe5641.
- Bramley, N., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, 124(3), 301.
- Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, 1(11), 769–771.
- Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, 127(3), 412.
- De Martino, B., O'Doherty, J. P., Ray, D., Bossaerts, P., & Camerer, C. (2013). In the mind of the market: Theory of mind biases value computation during financial bubbles. *Neuron*, 79(6), 1222–1231.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998–998.
- Fränken, J.-P., & Pilditch, T. (2021). Cascades across networks are sufficient for the formation of echo chambers: An agent-based model. *Journal of Artificial Societies and Social Simulation*, 24(3).
- Fränken, J.-P., Theodoropoulos, N. C., & Bramley, N. R. (2022). Algorithms of adaptation in inductive inference. *Cognitive Psychology*, 137, 101506.
- Fränken, J.-P., Theodoropoulos, N. C., Moore, A. B., & Bramley, N. R. (2020). Belief revision in a micro-social network: Modeling sensitivity to statistical dependencies in social learning. In *Proceedings of the 42nd annual meeting of the cognitive science society*.
- Fränken, J.-P., Valentin, S., Lucas, C., & Bramley, N. R. (2021). Know your network: Sensitivity to structure in social learning. In *Proceedings of the 43rd annual meeting of the cognitive science society* (Vol. 43).
- Hahn, U., Hansen, J. U., & Olsson, E. J. (2020). Truth tracking performance of social networks: How connectivity and clustering can make groups less competent. *Synthese*, 197(4), 1511–1541.
- Hahn, U., Harris, A. J., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, 29(4), 337–367.
- Harris, A. J., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: quantitative support for a bayesian network approach. *Cognitive Science*, 40(6), 1496–1533.
- Hawthorne-Madell, D., & Goodman, N. D. (2019). Reasoning about social sources to learn from actions and outcomes. *Decision*, 6(1), 17.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1), 1–55.
- Jasny, L., Waggle, J., & Fisher, D. R. (2015). An empirical examination of echo chambers in us climate policy networks. *Nature Climate Change*, 5(8), 782–786.
- Levin, S. A., Milner, H. V., & Perrings, C. (2021). The dynamics of political polarization.

- Proceedings of the National Academy of Sciences*, 118(50), e2116950118.
- Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3), 106–131.
- Lewandowsky, S., Pilditch, T. D., Madsen, J. K., Oreskes, N., & Risbey, J. S. (2019). Influence and seepage: An evidence-resistant minority can affect public opinion and scientific belief formation. *Cognition*, 188, 124–139.
- Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, 25(1), 322–349.
- Madsen, J. K., Bailey, R. M., & Pilditch, T. D. (2018). Large networks of rational agents form persistent echo chambers. *Scientific Reports*, 8(1), 12391.
- Madsen, J. K., Hahn, U., & Pilditch, T. D. (2020). The impact of partial source dependence on belief and reliability revision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Madsen, J. K., & Pilditch, T. D. (2018). A method for evaluating cognitively informed micro-targeted campaign strategies: An agent-based model proof of principle. *PloS One*, 13(4), e0193909.
- Scheufele, D. A., Hoffman, A. J., Neeley, L., & Reid, C. M. (2021). Misinformation about science in the public sphere. *Proceedings of the National Academy of Sciences*, 118(15), e2104068118.
- Tokita, C. K., Guess, A. M., & Tarnita, C. E. (2021). Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences*, 118(50).
- Whalen, A., Griffiths, T. L., & Buchsbaum, D. (2018). Sensitivity to shared information in social learning. *Cognitive Science*, 42(1), 168–187.
- Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2), 1.