Naïve Information Aggregation in Human Social Learning

J.-Philipp Fränken
Stanford University

Simon Valentin, Christopher G. Lucas, Neil R. Bramley
The University of Edinburgh

August 4, 2023

To glean accurate information from social networks, people should distinguish evidence from hearsay. For example, when testimony depends on others' beliefs as much as on first-hand information, there is a danger of evidence becoming inflated or ignored as it passes from person to person. We compare human inferences with an idealized rational account that anticipates and adjusts for these dependencies by evaluating peers' communications with respect to the underlying communication pathways. We report on three multi-player experiments examining the dynamics of both mixed human–artificial and all-human social networks. Our analyses suggest that most human inferences are best described by a naïve learning account that is insensitive to known or inferred dependencies between network peers. Consequently, we find that simulated social learners that assume their peers behave rationally make systematic judgement errors when reasoning on the basis of actual human communications. We suggest human groups learn collectively through naïve signalling and aggregation that is computationally efficient and surprisingly robust. Overall, our results challenge the idea that everyday social inference is well captured by idealized rational accounts and provide insight into the conditions under which collective wisdom can emerge from social interactions.

keywords: social learning; testimony; causal inference; Bayesian modeling

---

Word count: 7456

Contents

Introduction

Social learning is a key driver of the success of the human species (Henrich, 2017). It enables people to acquire knowledge vicariously, by observing the behavior of others (Bandura & McClelland, 1977), and via more explicit forms of information exchange (Jern & Kemp, 2015; Lucas et al., 2014). When it works, social learning can be advantageous (Laland, 2004; Rendell et al., 2011), supporting efficient learning and rational behavior at a population level (Krafft, Shmueli, Griffiths, Tenenbaum, & Pentland, 2020)—thereby enabling transmission of accumulated insights and wisdom to bootstrap future generations' learning (Kleiman-Weiner, Sosa, Gershman, & Cushman, 2019). When it falls short, social learning can lead to poor collective outcomes. For instance, repeated sharing of redundant information between agents can results 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).

One important aspect of successful social learning is the ability to distinguish the communication of new first-hand evidence from hearsay (Berg, 1993; Budescu & Yu, 2007; Enke & Zimmermann, 2019; Hahn, Hansen, & Olsson, 2020; Jönsson, Hahn, & Olsson, 2015; Whalen, Griffiths, & Buchsbaum, 2018). When agents communicate with one another, or base their beliefs on shared evidence, there is a danger of information becoming inflated or ignored as it spreads through a social network. For example, when two colleagues B and C recommend a new local coffee shop, it might seem like a ringing endorsement, until you learn that only B has been to the coffee shop while C just heard about it from B. This makes C's testimony non-independent and in this case, practically worthless. From the perspective of rational analysis (Oaksford, Chater, et al., 2007), dependencies between agents determine how much weight one of them should place on what each of the others says. Specifically, rational social learners should use their knowledge about who tends to communicate with whom—the structure of their social network—to make inferences about the truth behind a claim. Similarly, rational social learners should consider the timing and content of agents' statements to infer who is learning from whom, and when social judgments are indicative of new first-hand evidence.

In line with these predictions, recent empirical studies have shown correlations between human inferences and rational model simulations when reasoning with dependent sources of information (Fränken, Theodoropoulos, Moore, & Bramley, 2020; Whalen et al., 2018) as well as concordance between a rational theory-of-mind based framework and problems of reverse engineering the intended meaning of others' utterances or the beliefs and desires that caused their behavior more generally (Goodman & Stuhlmüller, 2013; Hawthorne-Madell & Goodman, 2019; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015; Lucas et al., 2014; Wu et al., 2021). In these works, social inference is grounded in a "utility-maximizing" assumption, making it possible to reverse engineer the causes of peers' behavior by inverting a generative model of how they ought to rationally update and express their beliefs to achieve their goals (Baker,

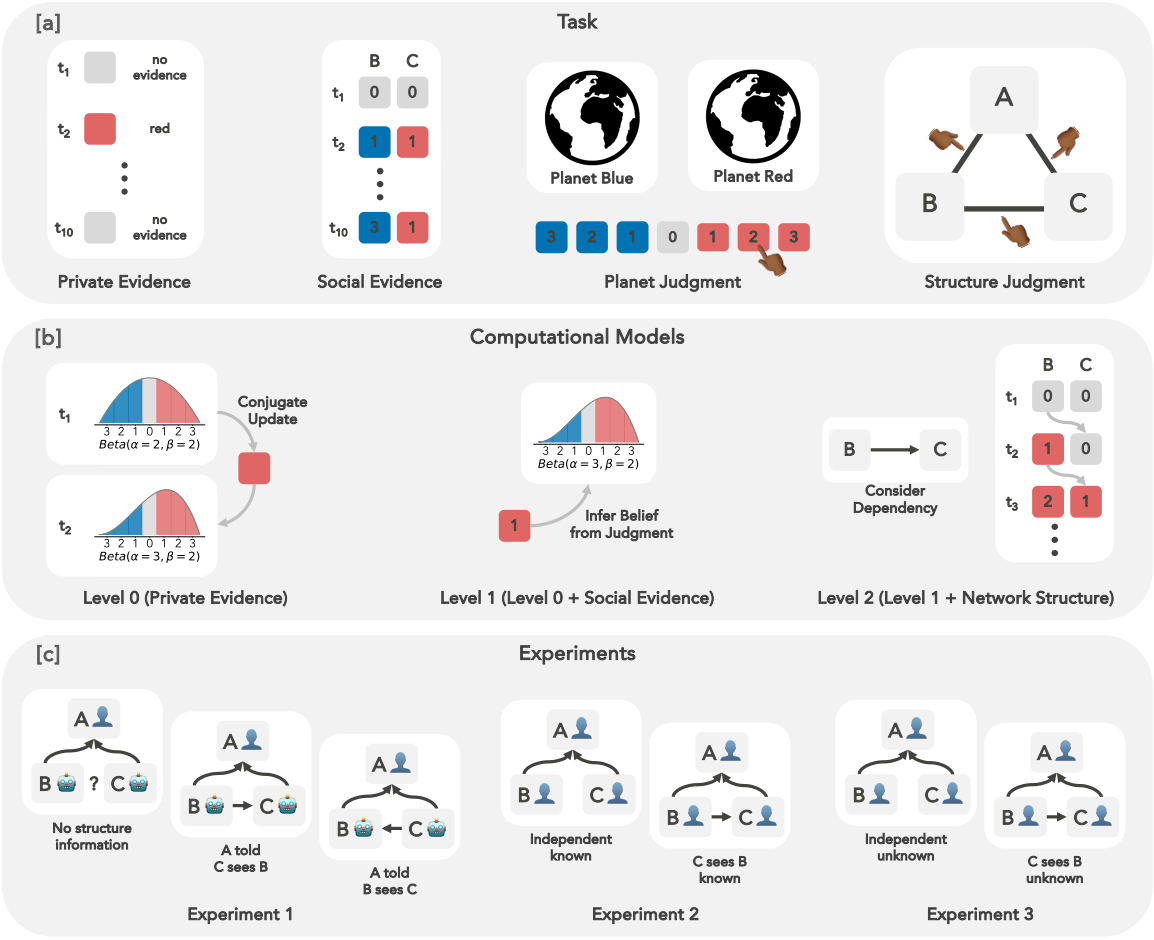

Figure 1. High-level overview of the paper. **[a]** Task illustration from the perspective of agent A (the focal participant across experiments). Over the course of ten trials, participants reason about their unknown location in space, i.e., whether they are on Planet Blue or Planet Red. Each planet has a different proportion of red and blue fish. On Planet Blue, $\frac{2}{3}$ of the fish are blue and $\frac{1}{3}$ are red. On Planet Red, the proportions are reversed. These proportions are known to participants. At each trial t, participants sample private evidence from the unknown planet $\in$ {blue fish: ■, red fish: ■, no evidence: ■} and observe social evidence corresponding to the previous judgments provided by two other agents (agents B and C). Participants then provide judgments about the planet they are on using a seven-point scale, ranging from 3–blue (highly confident planet blue) to 3–red (highly confident planet red), as well as a structure judgment about the communication structure between all three agents. Participants know that other agents' judgments are elicited using the same seven-point scale and that all are incentivised to be correct. **[b]** Illustration of computational models. The baseline model (Level-0) uses a beta-binomial model to update its beliefs strictly from private evidence. The naïve social learning model (Level-1) aggregates other agents' judgments and inverts a generative model of the observed judgments to infer the underlying belief (and thereby the evidence), followed by beta-binomial updates combining both observed private evidence and other agents' inferred private evidence. The idealized rational model (Level-2) further conditions its inferences about the unknown private evidence on the communication structure (i.e., it corrects for dependencies in evidence), followed by the same beta-binomial updates as the Level-1 model. **[c]** Experimental setup. In Experiment 1, participants (agent A) interact with two idealized (artificial) agents whose judgments were fixed across conditions. Between conditions, we manipulate whether participants (agent A) receive no structure information, are told that C could see B's judgments or the reverse. In Experiments 2–3, the focal participant (agent A) interacts with two other human participants (B & C) and the actual network structure is manipulated across two conditions. Participants are informed about the network structure in Experiment 2 but must infer it in Experiment 3.

Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Jara-Ettinger, Schulz, & Tenenbaum, 2020; Lopez-Brau, Kwon, & Jara-Ettinger, 2022).

While there is good evidence people are capable of these kinds of inferences, in this paper, we challenge the idea that social learning is, *in general*, well captured by such computationally involved and idealized rational accounts. To investigate this, we study human learning in micro social networks made up of both simulated and actual human learners. We find instead that social inference in this iterated online multi-person setting is more often dominated by computationally cheap "naïve" inferencing in which learners integrate their peers judgments at face value, lacking sensitivity to the nuances of their dependencies and redundancies. Inspired by standard "urn" tasks (Anderson & Holt, 1997), we develop a novel social learning task in which participants combine first-hand observations (*private evidence*) with judgments from two other agents who are making their own private observations (*social evidence*). Participants are tasked with inferring a property of their shared environment (*planet judgment*) and in some conditions also with inferring the underlying communication structure between themselves and the other agents (*structure judgment*; Fig. 1a). Since participants cannot directly access each other's private evidence, to succeed they must estimate the summative impact of the unknown private evidence seen by the other agents (simulated or human) from observations of the their public judgments and combine this with their own private evidence. One way to approach this task is by attempting to invert a generative model of the origins of peers' judgments, taking into account the likely dependencies in the social evidence due to the structure of the network and iterated nature of the task (e.g., Whalen et al., 2018). In contrast, pilot work in a similar task (Fränken, Valentin, Lucas, & Bramley, 2021) suggested that this beahvior may be more of an exception than a norm: many participants may adopt a more naïve approach, where they simply incorporate other agents' judgments in a heuristic manner.

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: (1) A baseline model (Level-0) which ignores social evidence and forms inferences strictly from private evidence, (2) a computationally cheap, naïve social learning account (Level-1), based on Fränken et al. (2021), which aggregates other agents' judgments at their face value, and (3) an idealized rational account (Level-2) which, similar to rational models of testimony (e.g., Fränken et al., 2020; Pilditch, Hahn, Fenton, & Lagnado, 2020; Whalen et al., 2018; Xie & Hayes, 2022), conditions its inferences about the unknown evidence on the underlying communication structure between agents. Using our task and modeling infrastructure, we study human inferences across three behavioral experiments. Our experiments differ in terms of the number of human versus artificial agents making up the network, as well as in terms of these agents' knowledge about the structure of the network (Fig. 1c). In Experiment 1, there is just one human participant per network (agent A) who interacts with two artificial agents (B and C). In Experiments 2–3, a focal participant (agent A) interacts with two other human participants (agents B and C).

The rest of our paper is structured as follows: We first introduce our social learning task (Task), then formalize inferences in our task through the lens of our computational models (Computational Models). Next, we describe our three behavioral experiments and report our findings (Experiments). We then conclude with a discussion of the broader

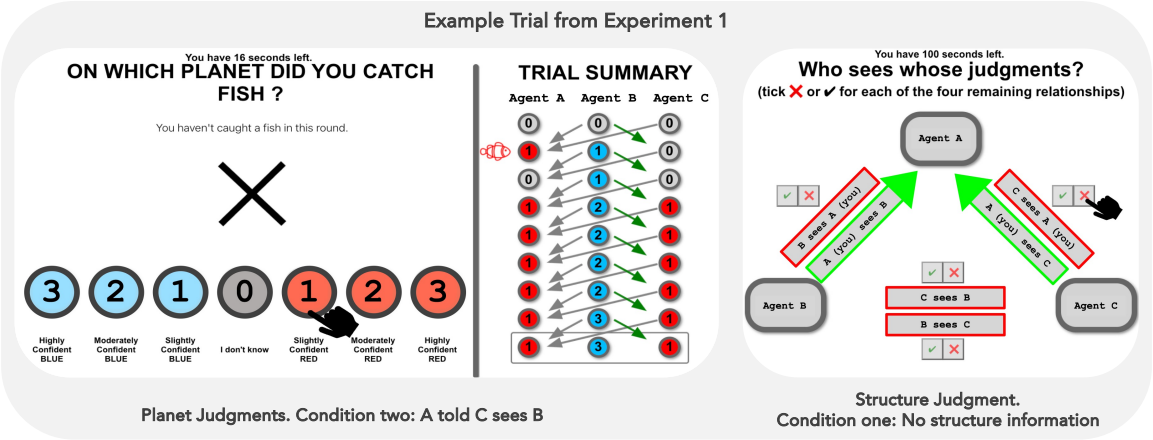

Figure 2. Example trial from Experiment 1. **Left**: Participants (agent A) provide a planet judgment at trial ten. At this trial, the participant did not observe additional private evidence (i.e., they did not catch fish). Previous private evidence (one red fish ■ observed at trial two), the participant's own judgments, as well as the previous judgments provided by agents B and C are shown in the trial summary section. In the displayed condition (A told C sees B), the communication structure was known to participants. In addition to explaining the communication structure to participants in the instructions (Fig. A-1), the trial summary section includes arrows to indicate who can see whose judgments. **Right**: In condition one with no structure information, participants (agent A) provide a structure judgment after completing all ten planet judgments.

impact and also the limitations of our work (General Discussion).

Task

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.

Alongside their own private observations, participants can see the previous judgments provided by two other agents. 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.

When inferring the unknown private evidence from other agents' judgments, participants may or may not be told about the communication structure between themselves and the other two agents. 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.

Computational Models

Overview We formalize learning in our tasks through the lens of Bayesian inference. At a high-level, the idea is that agents assign different degrees of belief to different states of the world, which we represent using probability distributions. As they encounter new private (binomial) evidence $\in$ {blue fish: ■, red fish: ■, no evidence: ■} at each trial, agents can update this distribution using conjugate updates within a beta-binomial scheme (■ $\rightarrow$ no update; see Fig. 3, left). Given a belief such as $h \sim Beta(\alpha = 3, \beta = 2)$, we assume agents derive probability masses for each of the $k = 7$ discrete judgment options using the cumulative density of the beta distribution $I_x(\alpha, \beta)$:

$$p(y_i \mid h) = \int_{(y_i-1)/k}^{y_i/k} \text{Beta}(x; \alpha, \beta) dx \tag{1}$$

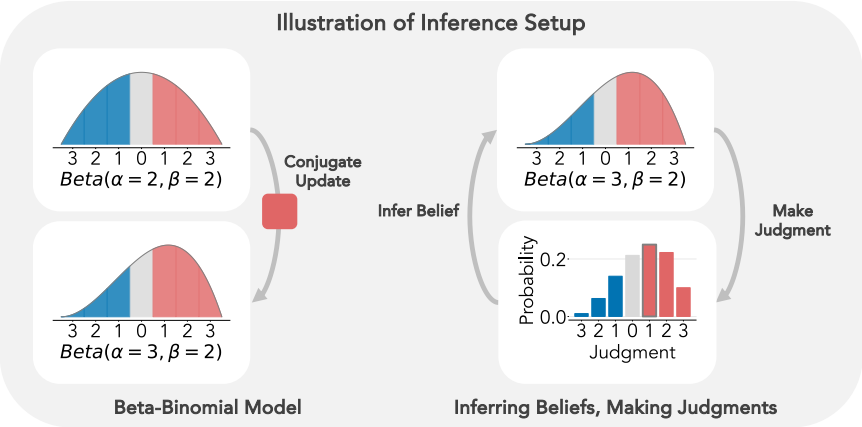

Figure 3. Illustration of inference setup. **Left**: agents update their beliefs upon observing private evidence via conjugate updates to their beta distribution. For example, upon observing a red fish, an agent increments the α parameter by one to update its belief from $h \sim Beta(\alpha = 2, \beta = 2)$ to $h \sim Beta(\alpha = 3, \beta = 2)$. **Right**: agents use their beliefs $h \sim Beta(\alpha, \beta)$ to make judgments by assigning probabilities to each response option on the seven-point scale. Similarly, given an observed judgment, such as "1-red" which corresponds to the highest bar with a grey border, agents can infer the most likely belief (and thus, the underlying evidence) that must have produced the observed judgment by inverting the model. This setup allows agents to flexibly infer unknown private evidence from observed judgments.

for each y_i where $i \in \{1, 2, ..., k\}$ (see Fig. 3, right). Moreover, given an observed judgment, such as "1–red", shown in Fig. 3, one can infer a distribution over possible beliefs h and thus make a guess about the evidence going into this judgment by "inverting" the inference model. The primary difference between our computational models listed below lies in the mechanism used to apply this inversion. This allows us to study whether people are primarily using computationally cheap, naïve (Level-1) inferences—which were previously explored in a simplified setting in Fränken et al. (2021)—or whether people use rational, structure-sensitive (Level-2) inferences—as proposed in other previous work (e.g., Fränken et al., 2020; Pilditch et al., 2020; Whalen et al., 2018).

Baseline: Level-0 Inference Our baseline model, Level-0 inference, disregards social evidence and thus applies no inversion of other agents' beliefs. Instead, the Level-0 model strictly updates beliefs $h \sim Beta(\alpha, \beta)$ upon observing private evidence by sequential application of Bayes' rule $p(h \mid d_t) \propto p(d_t \mid h)p(h)$ at each trial t using simple conjugate updates:

$$\alpha_t = \begin{cases} \alpha_{t-1} + 1 \text{ if } d_t = \blacksquare \\ \alpha_{t-1} \end{cases} \qquad \beta_t = \begin{cases} \beta_{t-1} + 1 \text{ if } d_t = \blacksquare \\ \beta_{t-1} \end{cases} \tag{2}$$

Naïve Social Learning: Level-1 Inference Our first social inference model, Level-1 inference, combines observed private evidence with observed judgments from other agents to update beliefs $h \sim Beta(\alpha, \beta)$. Specifically, for every other agent i, Level-1 uses that agent's most recent judgment $y^{(i)}_{t-1}$ to compute a posterior probability for each of the agent's most recent private evidence observations $p(d^{(i)}_{t-1} \mid h^{(i)}_{t-2}, y^{(i)}_{t-1})$ given the agent's previous belief $h^{(i)}_{t-2}$ and observed judgment $y^{(i)}_{t-1}$:

$$p(d_{t-1}^{(i)} \mid h_{t-2}^{(i)}, y_{t-1}^{(i)}) \propto p(y_{t-1}^{(i)} \mid h_{t-2}^{(i)}, d_{t-1}^{(i)}) p(d_{t-1}^{(i)}) \tag{3}$$

Here, $h_{t-2}^{(i)}$ corresponds to agent i's own beta-distributed belief prior to incorporating their private evidence observation $d_{t-1}^{(i)}$. For simplicity, Level-1 inference assumes that other agents are "reliable" (c.f. Hawthorne-Madell & Goodman, 2019), meaning that they assign a higher likelihood to response bins with a high posterior probability under their beliefs (see Additional Model Details, for further details). Consequently, Level-1 inference produces a prediction about the most likely evidence $d_{t-1}^{(i)}$ leading to the observed judgment $y_{t-1}^{(i)}$. This inference captures how the previous belief $h_{t-2}^{(i)}$ of an agent i must have changed in response to the inferred evidence to produce the observed judgment. As an example, imagine that $h_{t-2}^{(i)}$ is equal to $Beta(\alpha = 1, \beta = 1)$. If we now observe a new judgment $y_{t-1}^{(i)}$ which is equal to to "1-blue", we would identify $d_{t-1}^{(i)} =$ ■ as the most probable evidence leading to the new judgment. Importantly, this inference process allows us to incorporate uncertainty, meaning that we update our belief about another agent's parameters ($\alpha = 1$ and $\beta = 1$) based on the probability of ■ versus ■ given $y_{t-1}^{(i)}$. For example, assuming a uniform prior over private evidence and a judgment such as "1-blue", we may arrive at a posterior probability of 0.95 for $d_{t-1}^{(i)} =$ ■, a probability of 0.01 for $d_{t-1}^{(i)} =$ ■, and a probability of 0.04 for $d_{t-1}^{(i)} =$ ■, meaning that the updated marginal belief about i's belief $h_{t-1}^{(i)}$ corresponds to $Beta(\alpha = 1.01, \beta = 1.95)$.

Level-1 inference assumes that all agents form beliefs independently of one another, that is, implicitly and erroneously assuming agents never observe each other's judgments. Given this assumption, Level-1 inferencers can update their belief h at each time step by conditioning jointly on their current private evidence d_t and inferred evidence behind the judgments $\in \{h_{t-1}^{(1)}, ..., h_{t-1}^{(n)}\}$ from n independent agents:

$$\begin{aligned} p(h \mid d_t, h_{t-1}^{(1)}, ..., h_{t-1}^{(n)}) \propto \\ p(h_{t-1}^{(1)}, ..., h_{t-1}^{(n)} \mid d_{t-1}^{(1)}, ..., d_{t-1}^{(n)}) p(d_t \mid h) p(h). \end{aligned} \tag{4}$$

where $p(h \mid d_t, h_{t-1}^{(1)}, ..., h_{t-1}^{(n)})$ corresponds to Level-1's posterior belief and $p(h_{t-1}^{(1)}, ..., h_{t-1}^{(n)} \mid d_{t-1}^{(1)}, ..., d_{t-1}^{(n)})$ are the inferred beliefs for n agents based on their own (unobserved) private evidence from the previous time step obtained from Equation 3.[1]

Rational Social Learning: Level-2 Inference Our most sophisticated model, Level-2 inference, extends Level-1 by considering whether agents can see each other's judgments during previous time steps. Level-2 inference corrects for such dependencies by evaluating agents' beliefs to build a joint probability distribution over the potential histories of private evidence observed by each agent which can then be marginalized over. If structure is unknown this can be done separately under every structure hypothesis $g \in G$ and the learner can additionally marginalize over their structure uncertainty. Formally, Level-2 inference can be expressed as:

[1] In our task, agents cannot see the judgments being made by other agents on the current trial t *prior* to providing their own judgments, which is why the indices in Equation 3 and Equation 4 refer to the other agents' judgments provided at the previous time step $t - 1$.

$$p(h_{t-1}^{(1)}, ..., h_{t-1}^{(n)} \mid d_{t-1}^{(1)}, ..., d_{t-1}^{(n)}, g) = \sum_{g \in G} p(h_{t-2}^{(1)}) \prod_{i=2}^{n} p(h_{t-1}^{(i)} \mid h_{t-2}^{(1)}, ..., h_{t-2}^{(i-1)}, d_{t-1}^{(i)}, g) p(g). \tag{5}$$

Here, $p(g)$ corresponds to the probability of a given network structure $g \in G$ and $p(h_{t-1}^{(i)} \mid h_{t-2}^{(1)}, ..., h_{t-2}^{(i-1)}, d_{t-1}^{(i)}, g)$ is computed recursively until a termination condition is met. In our setup, this termination condition could be (1) finding an agent that has no parents (i.e., an independent agent) or (2) the start of the game if there are no independent agents.

Importantly, in the present analyses, g was provided to Level-2 inference (either through experimental instructions or by eliciting structure judgments from participants), which allows us to disregard the marginalization over structures G. We describe how to update $P(G)$ given a judgment sequence in Appendix Structure Learning. To understand the intuition behind Level-2 inference, consider condition two (A told C sees B) from Experiment 1. Here, B is an independent agent and the parent of C, who can see B's judgments. To infer the evidence behind both B's and C's observed judgments, Level-2 inference involves first computing $p(d_{t-1}^{(B)} \mid h_{t-2}^{(B)}, y_{t-1}^{(B)})$ and updating $h_{t-1}^{(B)}$ based on the inferred evidence $d_{t-1}^{(B)}$. Next, it involves inferring $p(d_{t-1}^{(C)} \mid h_{t-2}^{(C)}, y_{t-1}^{(C)}, d_{t-2}^{(B)})$ which conditions on the putative evidence B observed two time steps ago $d_{t-2}^{(B)}$ which C presumably inferred and incorporated at the previous time step. Thereafter the model updates $h_{t-1}^{(C)}$ based on both the private evidence imputed for C $d_{t-1}^{(C)}$ and the private evidence C previously imputed from B $d_{t-2}^{(B)}$. Consequently, if for example C provides a judgment $y_{t-1}^{(C)}$ equal to "1–red" upon observing a judgment $y_{t-2}^{(B)}$ equal to "1–red" from B at $t-2$, Level-2 inference accounts for this dependency when inferring $d_{t-1}^{(C)}$, such that the evidence most likely to have produced the observed judgments is not counted twice (see dependency in Fig. 1b).[2]

Accommodating Autocorrelation: "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 $y_i \in \{y_1, y_2, ..., y_k\}$ followed by a softmax function, the second family of models incorporates an additional free parameter (mixture weight π), which mixes

[2] An example implementation of our algorithm is available in `agent.py` in our online repository.

each soft-maxed model prediction with the possibility of simply "sticking" to the previous judgment (see Additional Model Details, for 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.[3]

Experiments

Using our task and computational models, we study human inferences across three behavioral experiments. In Experiment 1, one participant (playing agent A) interacts with two simulated agents (agents B & C). In Experiments 2 and 3, a focal participant (agent A) interacts with two other real human participants (agents B & C). For each experiment, we first unpack participants' judgment patterns descriptively, followed by an aggregate (group-level) analysis comparing directional differences between conditions. We then evaluate each model's predictive accuracy on an individual participant level using leave-one-out cross-validation, which is the main focus of our analysis. 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.

Experiment 1

Setup. We first examined a controlled setting with one participant (agent A) and two artificial agents, B and C, each operating under known and unknown social network structures. Using this setup, we studied participants' inferences across three between-subject conditions. In the first condition, *no structure information*, participants did not receive any information about the relationship between B and C, thus requiring them to infer the underlying network structure at the end of the game. In the second condition, *A told C sees B*, and the third condition, *A told B sees C*, participants were instructed that agents B and C were non-independent changing the impact of their judgments from the Level-2 model perspective.

Participants. We recruited 150 adults from Prolific Academic (Palan & Schitter, 2018), aiming for a sample size of 50 participants randomly assigned to each condition. Four participants dropped out, resulting in a final sample of $N = 146$ (35.94 ± 15.52, 98 female, 48 male). Of these, 47 participants were assigned to the first condition (no structure information), 50 participants to the second condition (A told C sees B), and 49 participants to the third condition (A told B sees C). Participants received a payment equivalent to an hourly rate of £5.02. This payment included a base amount of £1.00 and a performance bonus of up to £1.50. To incentivise high-quality judgments, we paid each participant a small bonus of £0.15 on every trial in which the participant's judgment was in the correct direction on the scale from the midpoint, where the "correct" direction was determined by the combined amount of evidence observed by the entire network. Before starting the main task, participants completed a brief training to familiarize themselves with the game environment and reward structure. Detailed instructions are provided in Fig. A-1 and our online demo.

[3] "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.

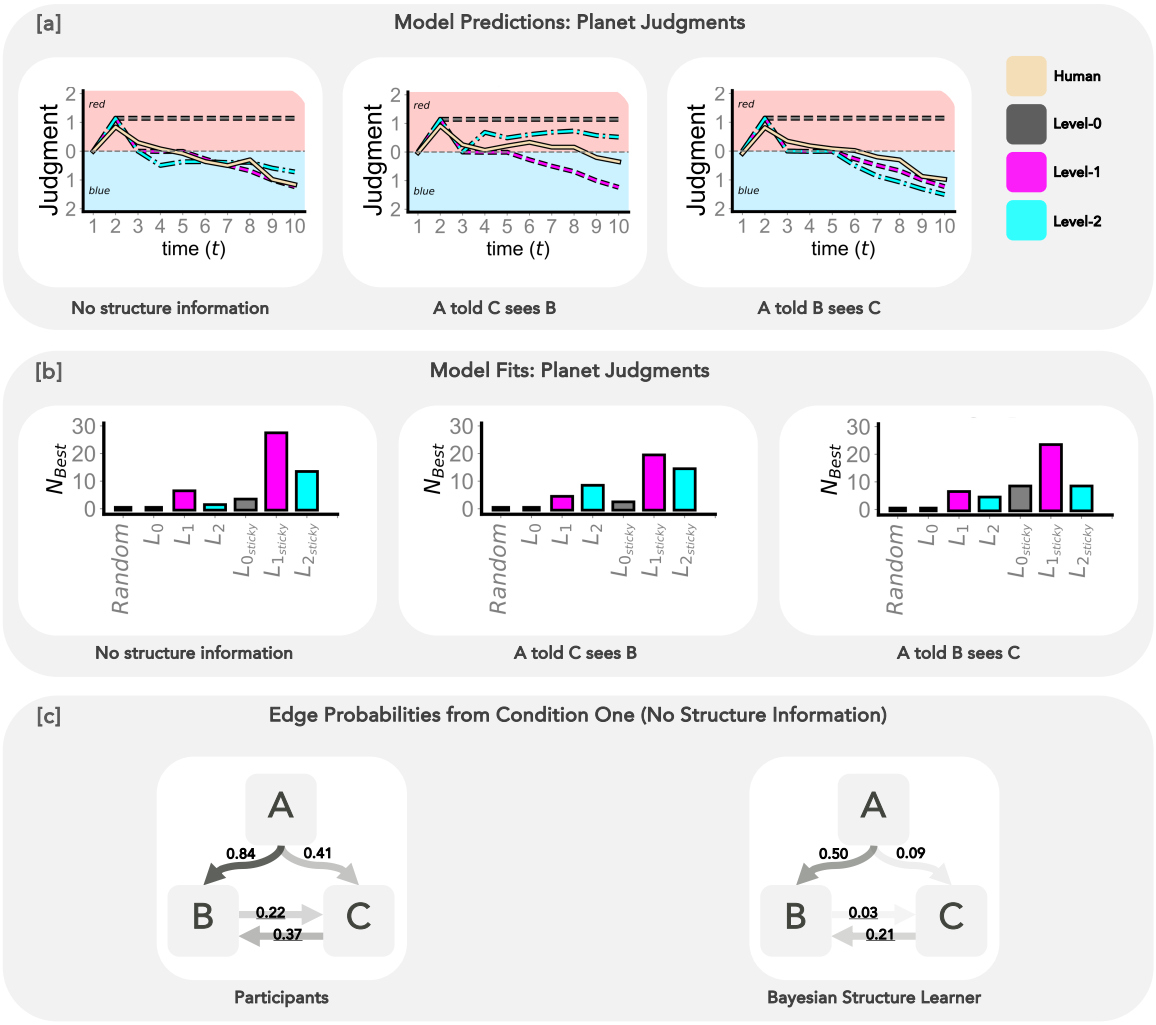

Figure 4. Results from Experiment 1. **[a]** Average human and model judgments (y-axis) across the ten time steps (x-axis) of our task for each condition. **[b]** Cross-validation results showing the number of participants best fit (y-axis) by each model (x-axis). **[c]** Average edge probabilities for participants (left) and a Bayesian structure learner (right; see Structure Learning) provided in condition one with no structure information.

Stimuli. 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.[4] 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

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

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.

Results. We begin by describing participants' average judgments (Fig. 4a) across different conditions. In the first condition (no structure information), the most likely private evidence distribution on the planet indicated participants were on Planet Blue. Consistent with this, 80% of participants gave a judgment in favor of Planet Blue at the final time step (mean ± standard error = 1.16 blue ± 0.20). In the second condition (A told C sees B), the most likely private distribution suggested participants were on Planet Red. However, only 34% of participants opted for Planet Red in the final judgment (and on average, preferred blue with a mean of 0.34 blue ± 0.19). Finally, in the third condition, 65% preferred blue at the final time step (0.98 blue ± 0.21), which again aligned with the true evidence distribution.

We next performed an aggregate group analysis to test for directional effects of our between-subject manipulation. Therefore, we averaged participants' judgments across time steps and performed a Kruskal-Wallis (Kruskal & Wallis, 1952) analysis of variance (ANOVA) which revealed a main effect of our manipulation (Kruskal's $H(2) = 13.73$, $p < 0.005$, $\epsilon^2 = 0.082$, 95%). Pairwise Mann-Whitney (Mann & Whitney, 1947) U-tests (one-sided) further confirmed this result: judgments in condition C sees B were significantly "redder" (mean judgment: 0.061 red, 95% confidence interval [0.152 blue, 0.274 red]) as compared to the first condition with no structure information (mean judgment: 0.471 blue, 95% confidence interval [0.763 blue, 0.179 blue], standardized U-score: $Z = 3.51$, $p < 0.001$, CLES = 0.707). Similarly, contrasting condition C sees B and condition B sees C (mean judgment: 0.297 blue, 95% confidence interval: [0.537 blue, 0.058 blue]) confirmed an effect of our structure manipulation, with a "redder" average judgment in condition C sees B compared to B sees C (standardized U-score: $Z = 2.36$, $p < 0.01$, CLES = 0.640). Both effects were significant at a Bonferroni-corrected significance level of $\frac{0.05}{3}$. There was no difference between condition one and three (standardized U-score: Z = -1.63, $p > 0.05$, CLES = 0.405).

To better understand individual-level behavior, we finally derived quantitative predictions for each inference model. The number of participants best predicted by each model during cross-validation within each condition is shown in Fig. 4b. Results from our individual-level analysis suggest predominantly naïve, structure-insensitive inferences with a bias towards sticking with the previous judgment (Level-1 sticky). Specifically, across all conditions, our Level-1 sticky account best predicted 48% of participants overall, while 24% across conditions were best characterized by the rational, structure-sensitive Level-2 sticky competitor. The majority of the remaining 28% across conditions were best accounted for by the Level-1 and Level-2 variants. These results suggest that a nontrivial number of participants were able to account for dependencies between agents' judgments in line with the predictions of Level-2 inference, while the majority were better described by a structure-insensitive naïve account.

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).

Overall, in Experiment 1, the dominant pattern of judgments reflected naïve social inference, with most participants overlooking the dependency between sources. This is also evidenced by a preference for blue across all three conditions at the final time step, despite the simulated agents' judgments being rationally consistent with there having been more red ■ than blue ■ fish caught overall in condition C sees B (see Fig. 4a). Our group-level analysis revealed that, averaged over time, participants' judgments in condition C sees B were significantly "redder" than in the other two conditions, qualitatively in line with the Level-2 account which best predicted 24% of participants' individual inferences across conditions. This suggests that, assuming a controlled setting with simulated agents, we are able to replicate previous structure-sensitive inference (e.g., Fränken et al., 2020; Whalen et al., 2018). However, our detailed investigation of individual-level behavior revealed that this statistical effect was driven by only a minority of the participants.

Experiment 2

Setup. A shortcoming of the simulated agent setting is that it is not clear how human-like the behaviour of the simulated agents is and what effect that has on the judgments. To explore a slightly more naturalistic setting, we next investigate inferences with three real participants. For Experiment 2, we focused on a known network structure setting and compared two between-subject conditions. In the first condition (*independent known*) participants playing the roles of B and C were independent of one another, that is they could not see each other's judgments but only received private evidence. In the second condition (*C sees B known*), participant C could see participant B's previous judgments, making C's judgments dependent on B. All participants were made aware of the communication structure prior to starting the task and it was visualized throughout the experiment (see Figure 2a). Apart from the above differences, instructions and procedures in Experiment 2 were identical to those in Experiment 1. Since this network structure means that Agent A is the only one that has to deal with dependent information sources, we report our primary analyses from the perspective of the focal participant (agent A).

Participants. To recruit participants via Prolific, we developed a client/server application enabling real-time interactions between three randomly matched participants using their web browser. Once matched, each triad was randomly assigned to one of the two between-subject conditions and the three participants were randomly assigned to one of the three roles. Overall, we recruited 126 participants (25.82 ± 7.08, 70 female, 56 male) comprising 42 triads. Twenty-one triads (63 participants) were assigned to condition one

(independent known) and another 21 triads (63 participants) were assigned to condition two (C sees B known). Participants were paid as in Experiment 1.

Stimuli. 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) 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.

Results. At the final time step, there was no statistical difference between A's judgment preferences across conditions (Fig. 5a). In condition one (independent known), 52.4% of participants assigned to A preferred Planet Red (mean ± standard error = 0.29 red ± 0.29). In condition two (C sees B known), the preference for red remained the same at 52.4%, although the mean was smaller at 0.14 red ± 0.44, which may be a result of the structure manipulation, despite there being no differences between conditions when averaging A's judgments across time steps (standardised U-score: $Z = 0.70$, $p > 0.05$, CLES = 0.564). Full judgment sequences are presented in Figure A-3).[5]

We next repeated our cross-validation analysis, which again supported predominantly naïve inference with a bias towards sticking close to earlier judgments (Level-1_{sticky}) in condition two (C sees B known), best predicting 62% of participants compared to only 10% best accounted for by structure sensitive Level-2_{sticky} (Fig. 5b). This was a notably smaller percentage compared to Experiment 1. Note that in condition one (independent known), Level-1 and Level-2 inference predictions coincide as the naïve aggregation of evidence is the same as assuming independence. Overall, results from Experiment 2 replicate the pattern of predominantly naïve inferences found in Experiment 1 but now do so in a genuinely social scenario.

[5] As expected, 91.5% (condition one) and 100% (condition two) of participants assigned to B preferred blue at the final time step (as all they saw was one blue fish), while 47.6% (condition one) and 28.6% (condition two) of participants assigned to C preferred red at the final time step, which was in line with the fact that C could see B's judgments in the second condition.

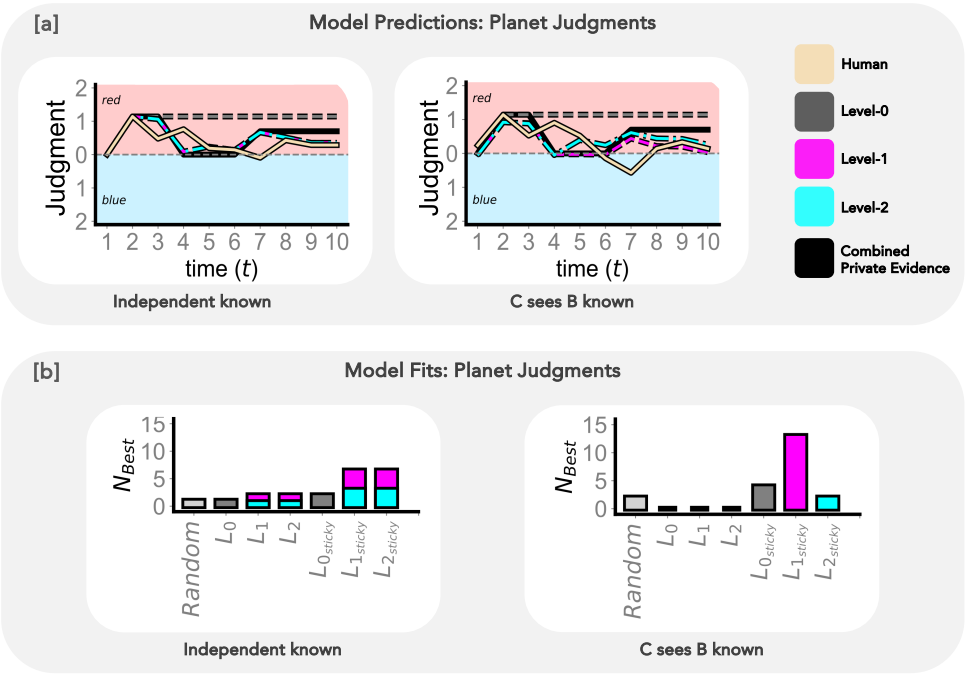

Figure 5. Results from Experiment 2. **[a]** Average human and model judgments (y-axis) across the ten time steps (x-axis) of our task for each condition. "Combined private evidence" refers to the ground truth, i.e., the predictions of an omniscient learner that had access to all private evidence observed by agents (see Fig. A-2b). **[b]** Cross-validation results showing the number of participants best fit (y-axis) by each model (x-axis).

Experiment 3

Setup. In Experiment 2, we gave participants complete access to the structure of their social network (i.e., full knowledge of who sees whose judgments). However, in real social interactions, we often have either no or only limited or uncertain knowledge about other people's precise communication histories. To assess how this additional layer of uncertaint and complexity affects inferences in the current learning problem, we finally studied a setting with three human agents and an initially unknown social network structure. As with the first condition from Experiment 1, participants had to provide structure judgments at the end of the task which we used for the Level-2 model.

Participants. We used the same client/server software as in Experiment 2 to sychnronize participants into triads. Overall, we recruited 129 adults (25.22 ± 7.04 years, 61 female, 68 male) through Prolific and paid as in Experiments 1–2. Procedures were identical to Experiment 2 with the only exception being the omission of the network structure instruction and structure visualization throughout the task. Overall, 22 triads (63 participants) were assigned to condition one (*independent unknown*) 21 triads (63 participants) were assigned to condition two (*C sees B unknown*).

Stimuli. As in Experiment 2, agent A caught a red fish at trial two, B caught a blue fish at trial three, and C caught a red fish on trial seven.

Results. There was no notable difference between A's judgment preferences across conditions at the final time step (Fig. 6a). In condition one (independent unknown), 45.5% of participants assigned to A preferred Planet Red (mean ± standard error = 0.41 red ±

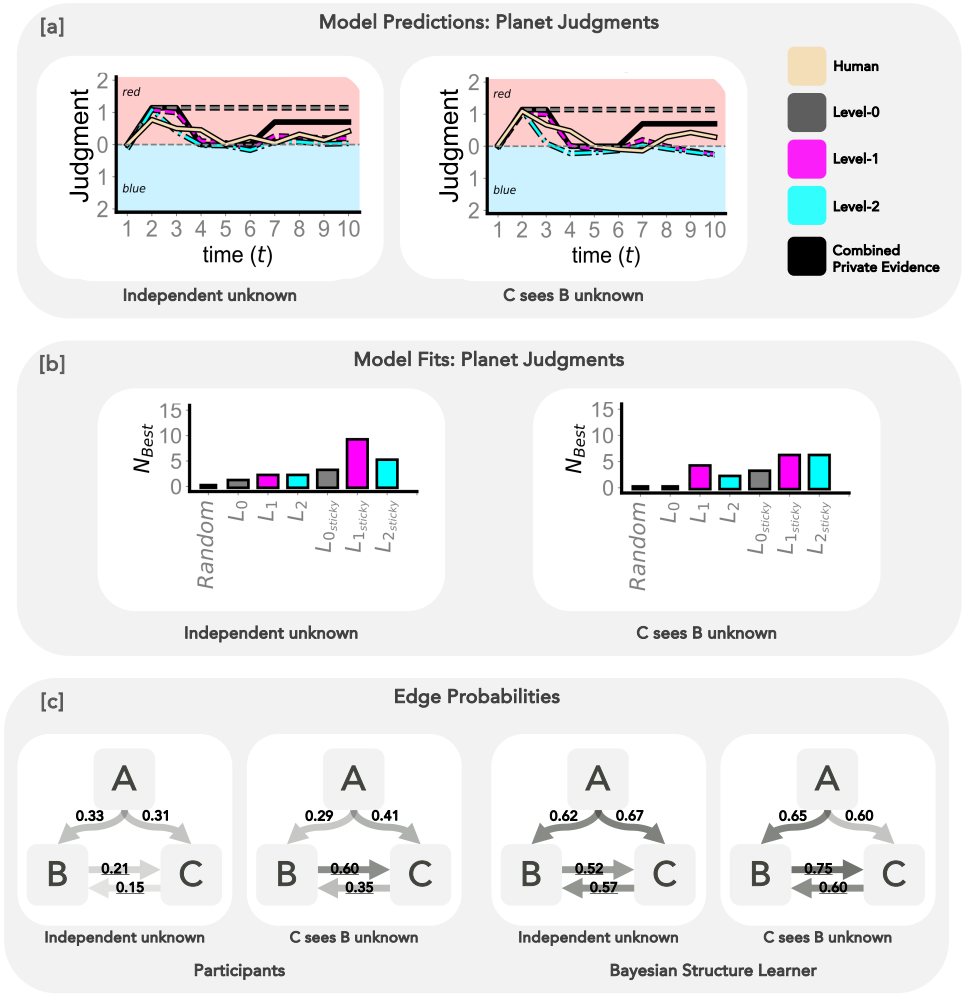

Figure 6. Results from Experiment 3. **[a]** Average human and model judgments (y-axis) across the ten time steps (x-axis) of our task for each condition. "Combined private evidence" refers to the ground truth, i.e., the predictions of an omniscient learner that had access to all private evidence observed by agents (see Fig. A-2c). **[b]** Cross-validation results showing the number of participants best fit (y-axis) by each model (x-axis). **[c]** Average edge probabilities for participants (left) and a Bayesian structure learner. Inferred edge probabilities from B to C in condition two were higher (60% for participants and 75% for a normative structure learner) as compared to condition one in which B and C were independent (31% for participants and 60% for a normative structure learner).

0.38). In condition two (C sees B known), the preference for red remained similar at 47.6%, again with a lower average judgment of 0.29 red ± 0.25 which was presumably a result of the structure manipulation, despite there being no significant difference between conditions when looking at A's time-averaged judgment sequence (standardised U-score: $Z = 0.486$, $p > 0.05$, CLES $= 0.544$).[6]

Our cross-validation model fit for Experiment 3 revealed that Level-1_{sticky} best pre-

[6] 100% (condition one) and 100% (condition two) of participants assigned to B preferred blue at the final time step (again, all they saw was one blue fish). 68.2% (condition one) and 19% (condition two) of participants assigned to C preferred red at the final time step, which was again in line with the fact that C could see B's judgments in the second condition.

dicted 41% of participants in condition one, which was again higher than Level-2_{sticky} (23%; see Fig. 6b). In condition two, Level-1_{Sticky} and Level-2_{Sticky} inference (conditioning on particpants' finally judged structure, see below) shared the same proportion of participants best predicted (29%). Overall, these results provide additional support for the previously found naïve pattern across conditions.

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 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.

Differences between Experiments

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.

General Discussion

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 et al., 2020; Lewandowsky, Pilditch, Madsen, Oreskes, & Risbey, 2019; Madsen, Bai-

ley, & 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.

Building on a distinct literature that has developed computational models of rational social learning, we noted how previous social inference accounts depend on a utility-maximizing assumption under which people reverse engineer one another's mental states and the evidence behind them by inverting a generative model of how those people form and express their beliefs. Our experiments probed the extent to which this framework captures real multi-agent social learning dynamics. The results suggest that while a minority of participants exhibit hallmarks of sensitivity to the principles of such rational social inferences, the dominant inference mode is more naïve, seeming to lack accommodation of communication-history-based dependencies between peers. Our results thus challenge a number of recent findings suggesting that human social learners reliably engage in sophisticated, rational inferences when reasoning from others' berhavior (Baker et al., 2017; Fränken et al., 2020; Whalen et al., 2018). 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.

Moreover, results from our all-human networks with noisy heterogeneous communication patterns demonstrate that assuming other agents are rational reasoners (e.g., Jara-Ettinger et al., 2015) can also be a cause of systematic judgment errors when those agents fall short of this standard (Fig. 6a). Contrary to the directionally wrong predictions by the rational Level-2 model predictions, human inferences were surprisingly robust with most agent A participants ending up with a directionally correct belief about the planet they were on. This might suggest we adopt simpler models of our peers than the utility-maximizing "homo economicus" central to more idealized accounts (Baker et al., 2017, 2009; Jara-Ettinger et al., 2020), thereby accounting for the fact that others' reasoning strategies are also often fallible and naïve. This is essentially the opposite move to attempting to accommodate such inter-individual limitations and heterogeneity explicitly in one's social reasoning. For instance, one could articulate a "Level-3" extension that attempts to anticipate the various departures peers make from rational inferencing (Alanqary et al., 2021). However, this approach is computationally expensive and demands additional theory-of-mind recursion and costly marginalization (Alon, Schulz, Dayan, & Rosenschein, 2022; Camerer, Ho, & Chong, 2004; Oey, Schachner, & Vul, 2022). On the face of it such complex reasoning seems unlikely to be resource rational in most circumstances (Lieder et al., 2018) given that the computational limitations of the social reasoner are on average going to be just as severe as those plaguing their social peers. As such, we propose that, collectively, human societies

may find a better computation-accuracy trade-off in the social inference sphere through mutual adoption of a more naïve social learning heuristics.

There are obvious limitations to our study: While our focus on human–human interactions extends previous work that was restricted to simulated social peers (Enke & Zimmermann, 2019; Fränken et al., 2020) and known dependencies (Pilditch et al., 2020; Whalen et al., 2018), the limited degree to which participants engaged in rational inferences might be a result of the low-bandwidth of the social evidence (i.e., rating scales) or the incentive structure (participants were rewarded for their own success irrespective of others). Future extensions of the present paradigm could thus explore the consequence of allowing participants to provide richer, more linguistic, social evidence—and so increase the bandwith of communication channels in the network. Moreover, extensions could examine the impact of cooperative incentives, which might lead participants to signal strategically and make correspondingly different social inferences in order to maximize a group's overall payoff. Such a setting might plausibly result in more deeply recursive social inferences, as a learner's reward would directly depend on the beliefs and performance of their peers. Cooperative incentives would also open the door to incorporating theories of active learning (Coenen, Nelson, & Gureckis, 2019) and metacognition (Fleming & Daw, 2017) to capture how people might use their communication signals to resolve uncertainty about peers' competence, motivations, or about the social network structure. 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. Furthermore, given that people's judgments in our task were best described by autocorrelated ("sticky") variants of our models, another extension could be to incorporate others' tendency towards autocorrelation when reasoning about their judgments. Finally, it is important to further establish the conditions under which naïve inference can be adaptive, as well as its implications for explaining population dynamics like information cascades (Bikhchandani, Hirshleifer, & Welch, 1992) and echo chambers (Madsen & Pilditch, 2018).

In sum, we found that people's inferences in an iterated social learning setting were relatively insensitive to dependencies between their network peers. Moreover, we found that simulated rational learners who assume peers behave rationally make systematic judgment errors when reasoning from genuine noisy human judgments. In contrast, human learners appear to succeed in our task through naïve inference strategies that are less sophisticated, but computationally efficient and surprisingly robust. We take this as suggestive that a comprehensive account of social cognition in the wild should take inferential naïvety seriously as feature of human social learning dynamics.

CRediT authorship contribution statement

J.-Philipp Fränken: Conceptualisation, Software, Formal analysis, Investigation, Data curation, Visualisation, Writing. **Simon Valentin**: Conceptualisation, Software, Formal Analysis, Writing. **Christopher G. Lucas**: Conceptualisation, Formal Analysis,

Supervision, Writing. **Neil R. Bramley**: Conceptualisation, Formal Analysis, Funding acquisition, Supervision, Writing.

Declaration of Interest

The authors declare no competing interest.

Acknowledgements

The work was supported by an EPSRC New Investigator Grant (EP/T033967/1) to Neil R. Bramley. J.-Philipp Fränken acknowledges support from the German Academic Scholarship Foundation. We would like to thank Lion Schulz, Alex Prodan, and Mark Gorenstein for useful comments on an earlier version of this draft.

Data Availability

The data and code supporting the findings in the present study are available on GitHub at janphilippfranken/spacefish.

Supplementary Information

A demo version of our task is available here.

References

Alanqary, A., Lin, G. Z., Le, J., Zhi-Xuan, T., Mansinghka, V. K., & Tenenbaum, J. B. (2021). Modeling the mistakes of boundedly rational agents within a bayesian theory of mind. *arXiv preprint arXiv:2106.13249*.

Alon, N., Schulz, L., Dayan, P., & Rosenschein, J. (2022). A (dis-) information theory of revealed and unrevealed preferences. In *Neurips 2022 workshop on information-theoretic principles in cognitive systems.*

Anderson, L. R., & Holt, C. A. (1997). Information cascades in the laboratory. *The American economic review*, 847–862.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bandura, A., & McClelland, D. C. (1977). *Social learning theory* (Vol. 1). Englewood cliffs Prentice Hall.

Berg, S. (1993). Condorcet's jury theorem, dependency among jurors. *Social Choice and Welfare*, *10*(1), 87–95.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, *100*(5), 992–1026.

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.

Budescu, D. V., & Yu, H.-T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*(2), 153–177.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861–898.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*(5), 1548–1587.

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.

Enke, B., & Zimmermann, F. (2019). Correlation neglect in belief formation. *The Review*

of Economic Studies, *86*(1), 313–332.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91.

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).

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

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.

Henrich, J. (2017). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

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.

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, *142*, 12–38.

Jönsson, M. L., Hahn, U., & Olsson, E. J. (2015). The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition*, *142*, 191–204.

Kleiman-Weiner, M., Sosa, F., Gershman, S., & Cushman, F. (2019). Downloading culture. zip: Social learning by program induction with execution traces. In *Cogsci* (p. 3495).

Krafft, P., Shmueli, E., Griffiths, T. L., Tenenbaum, J. B., & Pentland, A. (2020). Bayesian collective learning emerges from heuristic social learning. *Cognition*.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583–621.

Laland, K. N. (2004). Social learning strategies. *Animal Learning & Behavior*, *32*(1), 4–14.

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., 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.

Lopez-Brau, M., Kwon, J., & Jara-Ettinger, J. (2022). Social inferences from physical evidence via bayesian event reconstruction. *Journal of Experimental Psychology: General*.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS One*, *9*(3), e92160.

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.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Oaksford, M., Chater, N., et al. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oey, L. A., Schachner, A., & Vul, E. (2022). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*.

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Pilditch, T. D., Hahn, U., Fenton, N., & Lagnado, D. (2020). Dependencies in evidential reports: The case for informational advantages. *Cognition*, *204*, 104343.

Rendell, L., Boyd, R., Enquist, M., Feldman, M. W., Fogarty, L., & Laland, K. N. (2011). How copying affects the amount, evenness and persistence of cultural knowledge: insights from the social learning strategies tournament. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1118–1128.

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.

Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman-Weiner, M. (2021). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, *13*(2), 414–432.

Xie, B., & Hayes, B. (2022). Sensitivity to evidential dependencies in judgments under uncertainty. *Cognitive Science*, *46*(5), e13144.

Appendix

Additional Model Details. Following computation of discrete probabilities for each judgment $y_i \in \{y_1, y_2, ..., y_k\}$ given h, we computed the expected loss for each judgment:

$$L(y_i) = \sum_{\bar{y}} p(\bar{y} \mid h)(y_i - \bar{y})^2 \tag{A-1}$$

to penalize judgments further away from the mid-point of the scale prior to encountering evidence. As we assumed that participants started the game with a uniform prior

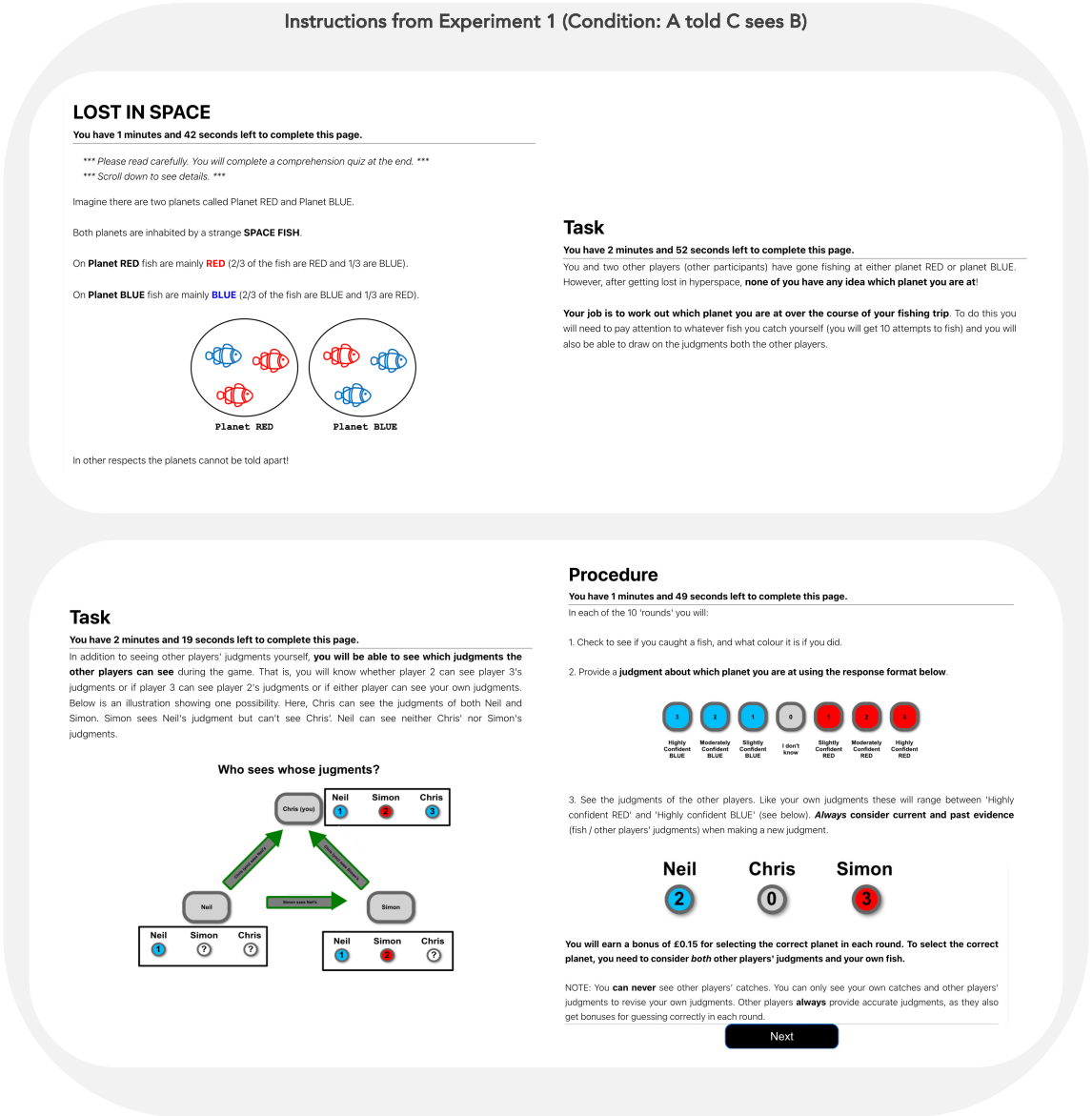

Figure A-1. Instructions shown to participants in Experiment 1 during condition two (A told C sees B). Full instructions available here.

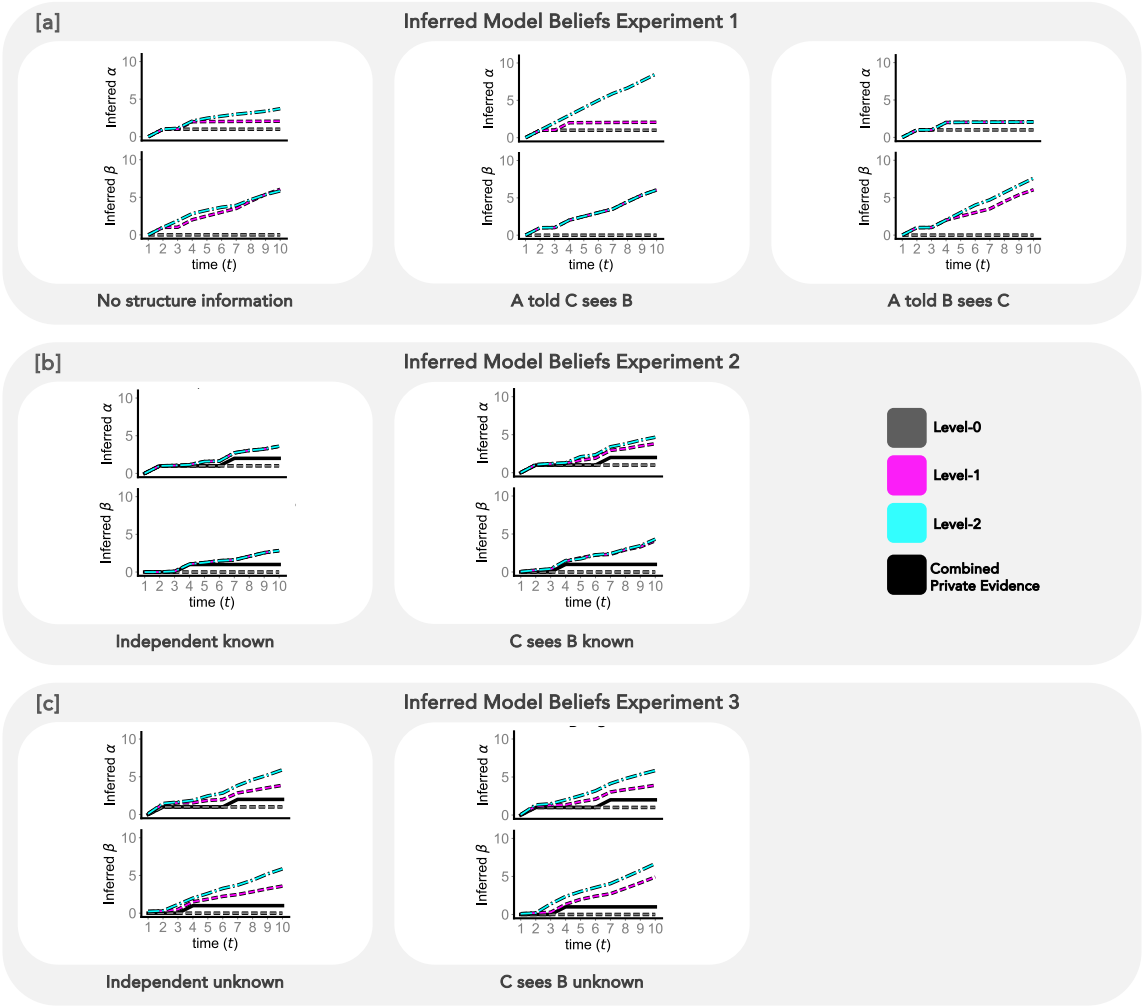

Figure A-2. Inferred beliefs for each inference model (y-axis) across time steps (x-axis) for each experiment and condition. α corresponds to the total number of observed + inferred red fish. β corresponds to the total number of observed + inferred blue fish. For Experiments 2–3, "combined private evidence" refers to the ground truth, i.e., the predictions of an omniscient learner who does not have to infer evidence from other agents' judgments but can "see" all the private evidence available to the network.

$Beta(1,1)$, this penalty effectively pushed model predictions towards "0" at the start of the game prior to observing private evidence or observing other agents' judgments.

We next used a softmax-function with inverse temperature parameter τ to obtain final choice probabilities for each $y_i \in \{y_1, y_2, ..., y_k\}$:

$$p(y_i \mid h) = \frac{e^{-L(y_i)\tau}}{\sum_{i=1}^{i=k} e^{-L(y_i)\tau}}. \tag{A-2}$$

We set $\tau = 10$ (high reliability), resulting in a strong (practically deterministic) preference for selecting the judgment with the highest probability given observations. For the sticky variant of our inference models, we included an additional mixture weight π to assess the weight that participants placed on their previous judgment $p(y_{t-1}^{(i)})$:

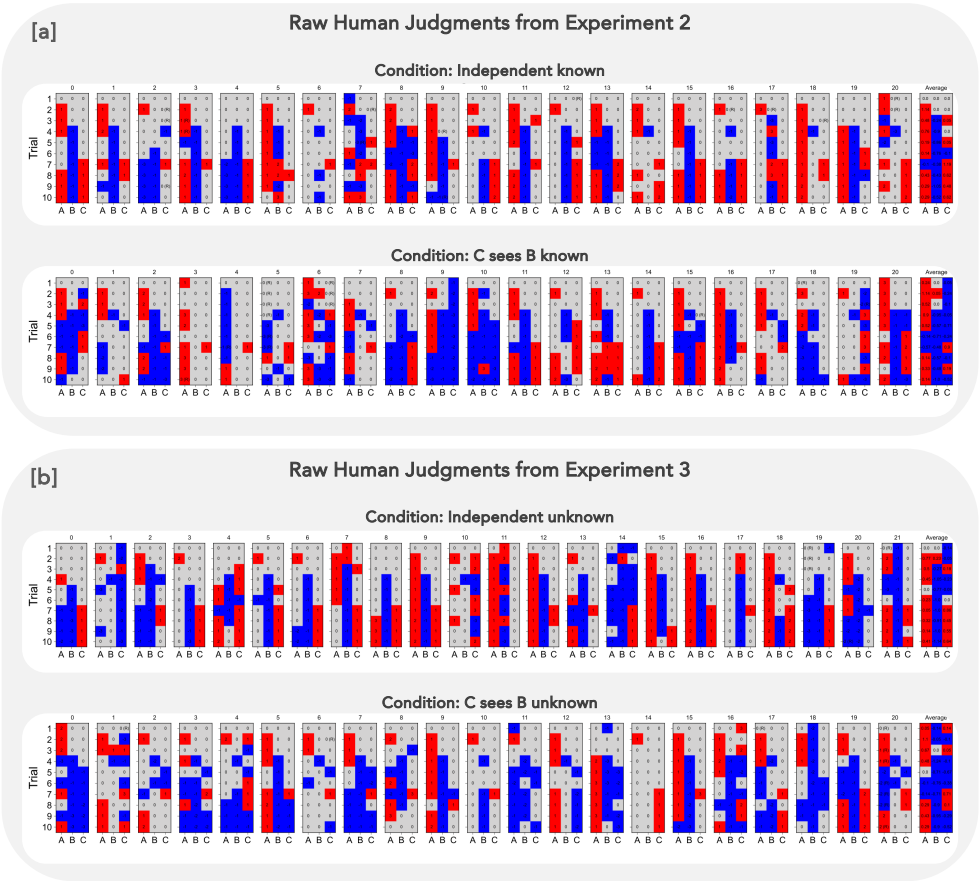

Figure A-3. **[a]** Raw judgments provided by participants in Experiment 2. **[b]** Raw judgments provided by participants in Experiment 3.

$$p(y_t^{(i)} \mid h) = \pi p(y_t^{(i)} \mid h) + (1-\pi)p(y_{t-1}^{(i)} \mid h). \tag{A-3}$$

We obtained values for π and τ by joint maximum likelihood optimisation using L-BFGS approximation implemented in SciPy.

Structure Learning. To sanity check our manipulations in the first condition of Experiment 1 as well as participants' ability to infer the unknown structures in Experiment 3, we used a Bayesian structure learning model to infer the most probable edges between agents from the observed communication sequences. Let's denote the set of judgments of all the agents at time t as $\mathbf{Y}_t = y_t^{(1)}, y_t^{(2)}, ..., y_t^{(n)}$, where $y_t^{(i)}$ is the judgment of agent i at time t. Let's denote $Y_{1:t}^{(i)}$ as the sequence of judgments for agent i from time 1 to time t.

We inferred the structure selections shown in Fig. 4 and Fig. 6 using a normative (Bayesian) change-based judgment approach with a single free parameter $\theta = 0.9$. This parameter represents the execution accuracy, that is, the probability with which the structure learner executes a deterministic policy to update their belief $P(G)$ about network structures.

The structure learner predicted a judgment change for the target agent v from time $t = -1$ to $t = 0$, conditional on potential changes in the judgments of agent v's parent agents $pa(v)$ from time $t = -2$ to $t = -1$. The structure learner combined changes in

evidence from multiple parent agents using an additive combination function:

$$\Delta_{Y_{t-2}^{pa(v)} \to Y_{t-1}^{pa(v)}} = \sum_{i \in pa(v)} \Delta_{y_{t-2}^{(i)} \to y_{t-1}^{(i)}}. \tag{A-4}$$

Given the assumptions above, the posterior probability of a specific network structure g underlying the sequences of judgments from n agents is given by:

$$p(g|Y_{1:t-1}^{(1)}, ..., Y_{1:t-1}^{(n)}) \propto p(Y_{1:t-1}^{(1)}, ..., Y_{1:t-1}^{(n)}|g)p(g), \tag{A-5}$$

where $p(g)$ is the prior probability of the structure (which is assumed to be uniform) and $p(Y_{1:t-1}^{(1)}, ..., Y_{1:t-1}^{(n)}|g)$ is the likelihood of the judgment sequences given g. For a given agent v and structure g, if $pa(v) =$ or $\Delta_{Y_{t-2}^{pa(v)} \to Y_{t-1}^{pa(v)}} = 0$, the likelihood that v keeps their previous state (i.e., $\Delta_{y_{t-1}^{(v)} \to y_{t0}^{(v)}} = 0$) is given by $\theta + \frac{1-\theta}{k}$ where $k = 7$ is the number of possible responses in our experiments. If $\Delta_{Y_{t-2}^{pa(v)} \to Y_{t-1}^{pa(v)}} > 0$, the likelihood of a non-negative change for v is given by $\frac{\theta}{k - y_{t-1}^{(v)} + 1} + \frac{1-\theta}{k}$. If $\Delta_{Y_{t-2}^{pa(v)} \to Y_{t-1}^{pa(v)}} < 0$, the likelihood of a non-positive change is given by $\frac{\theta}{y_{t-1}^{(v)}} + \frac{1-\theta}{k}$.