## Title
Agenda setting and The Emperor's New Clothes: people infer that letting powerful agents make their opinion known early can trigger information cascades and pluralistic ignorance



## Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

## Authors
Richardson, Emory
Davis, Isaac
Keil, Frank

## Publication Date
2023

Peer reviewed

# Agenda setting and *The Emperor's New Clothes*: people infer that letting powerful agents make their opinion known early can trigger information cascades and pluralistic ignorance

**Emory Richardson** (emory.richardson@yale.edu**), Isaac Davis** (isaac.davis@yale.edu), & **Frank Keil** (frank.keil@yale.edu)

**Abstract**

Consensus-based social learning strategies often outcompete other strategies in evolutionary models. But while formal proofs suggest that consensus' reliability is compromised when individual judgments are not independent, this makes for a notoriously implausible assumption in the biological world: the people we learn from are constantly learning from each other as well. How do we avoid being misled by consensus? We present two experiments and a computational model examining commonsense reasoning about how people's public and private judgments are influenced by the consensus and social status of those around them. Results suggest that while people realize that these two factors can cause others' public and private judgments to diverge, their own trust in public consensus depends on how accurately they believe it reflects their informants' true beliefs.

**Keywords:** collective behavior, agenda-setting, information

## Introduction

Imagine a five-person engineering team needing to decide which of two model airplane designs would fly the best in a contest. Each teammate first evaluates the airplanes privately and then, one-by-one, announces their vote. This means that only the first speaker votes without knowing anyone else's vote, and knowing earlier speakers' votes could change later speakers' public votes — and their private beliefs — for better or for worse. Moreover, it may not affect them in the same way, and might not affect them at all. As a truth-seeking fifth speaker, you might benefit from adjusting your confidence in your own answer after seeing your teammates' votes — but how could you distinguish the signal from the noise?

The inference problem you face is trickier than it is for earlier speakers: in order to weigh your own judgment against the previous speakers', you'll also need to evaluate speaker one's influence on two, their joint influence on speaker three, and so on. But it's not a problem easily dismissed as an artifact of an unusual voting process: many social contexts involve some form of sequential belief-updating at micro- and macro-scales, from turn-taking in conversation and parliamentary voting to viral tweets and the influence of yesterday's trades on today's stock prices. And it's a problem with teeth: your decision to trust or distrust that chain of influences not only affects you, but the downstream observers influenced by your decision, who may later influence you in turn. Socially transmitted information can be useful — cumulative culture would be impossible without it — but it can also lead to catastrophic information cascades (Boyd & Richerson, 1995; Raafat et al., 2009). So how do we learn from social information without being misled by it? Here, we focus on commonsense reasoning about how public and private judgments are influenced by *social favor* (the desire to align oneself with or against an individual or group) as well as *information cascades* (herd behavior by rational agents exposed to sequential decision-making processes). We examine this reasoning using human data from two experiments and a computational model.

Here's the gist of our argument. Large literatures demonstrate that stronger consensus can put dissenters under greater pressure to conform; and that social power (i.e., dominance, prestige) can have a similar influence (Raafat et al., 2009; Kameda et al., 2022; Morgan et al., 2012; Pink et al., 2021; Jiménez & Mesoudi, 2019). In both cases, this pressure can include epistemic motivations, but it can also come from simply wanting to align oneself with an individual or group — a desire for social favor, even against one's better judgment. But these pressures aren't abstruse discoveries of 20th century academic psychology. They're part of the commonsense psychology adults use to interpret each others' behavior (Gerstenberg & Tenenbaum, 2017; Baker et al., 2017). Our suggestion is that commonsense reasoning about how they play out in sequential updating processes can make people more vigilant to the potential for information cascades. But existing evidence of people's vigilance in sequential updating scenarios is mixed. We suggest that this is not because people are insensitive to the potential for information cascades. Instead, we'll argue below that understanding people's reliance on socially-transmitted information requires looking closely at how they expect it to affect other people's beliefs.

**(In)sensitivity to information cascades in sequential-updating.** Learning from Bob lets Alice avoid the costs of learning alone; but she also risks inheriting his errors. And the errors may not even be his. Bob may have learned from Carol, who learned from David, and so on. In other words, social learning allows *errors* to cascade through long transmission chains as well as knowledge. In some contexts (e.g., hearsay), the risks of relying on second- and third-hand information seem intuitively obvious (Altay, Claidière, & Mercier, 2020). But less intuitively, a few relatively mild assumptions make deferring rational in some contexts. Namely: if Alice assumes Bob's actions reflect the evidence available to him, then even a relatively small consensus may



In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.* 

provide her with sufficient evidence-of-evidence to override *whatever* her private evidence suggests. And evidence of evidence *is* evidence (Feldman, 2014; Dorst, 2022). For instance, suppose that given a binary choice, her private evidence suggests Bobcat Bite makes a better burger than Louis' Lunch. But she sees that Bob and Carol have chosen Louis' Lunch; if she assumes that these decisions reflect their (also binary) evidence, then her evidence consists of two votes for Louis' (Bob and Carol) and one for Bobcat (her). Being a rational agent, she changes her mind and goes to Louis'. The counter-intuitive implication noted in two seminal papers (Banerjee, 1992; Bikhchandani, Hirshleifer, & Welch, 1992) is that a relatively small number of votes are sufficient to ensure that every subsequent voter will face the same decision as Alice — but since they would conform regardless of their private evidence, their public decisions are, paradoxically, no longer informative to later voters. Rational deference allows the initial votes to cascade through a chain.

Given the theoretical importance of epistemic vigilance in social learning to theories of biological and cultural evolution, one might expect learners to be sensitive to the risks of information cascades, despite the counter-intuitiveness of the implication that "a rational Alice should always defer". But evidence from the several existing studies is mixed (A8derson & Holt, 1997; Whalen, Griffiths, & Buchsbaum, 2017; Xie & Hayes, 2022). Participants in these studies are presented with an urn mostly containing red balls or mostly blue. After privately sampling a ball from the urn, each informant either announces their inference publicly to the participant and remaining informants (potentially influencing their judgments), or privately tells the participant which color they believe predominates (ensuring each informant's judgment is independent). When every informant infers the same color, people were just as willing to defer to the unanimous consensus in the publicly announced sequence as the private sequence (Xie & Hayes, 2022). Moreover, even when experimenters showed participants both sequences and asked them to explain whether or not one would be more informative than the other, participants defended each in approximately equal proportions (~35-40%, with 25% indifferent). These findings are consistent with earlier work (Whalen, Griffiths, & Buchsbaum, 2018) in which participants' skepticism was only piqued if (A) one of three informants dissented or (B) all three informants jointly sampled a single ball to make their judgments instead of each sampling their own.

Why aren't people more skeptical of consensus when their informants could have been influenced by hearing each other's decisions? Like others (Xie & Hayes, 2022; Whalen, Griffiths, & Buchsbaum, 2018; Laan, Madirolas, & de Polavieja, 2017; Dietrich & Spiekermann, 2013), we think it's noteworthy that our informants' judgments are rarely independent in the real world. People are constantly observing each others' decisions, and allowing them to do so often makes consensus more reliable instead of less reliable (Kao et al., 2014; Barnett, 2019; Pilditch, Hahn, Fenton, Lagnado, 2020; Toyokawa, Whalen,and Laland, 2019). Moreover, they often rely *less* on others' judgments than models imply is optimal, and when they do defer, they're selective about who they trust and why (Mannes, 2009). So in real world contexts, one can't disregard consensus simply because informants *might* have influenced each other. One has to consider whether they *were* influenced, and whether that influence would make consensus more reliable or less. The balls-and-urn task guarantees that allowing informants to observe each others' decisions can only make consensus less reliable — but it may not be the kind of context in which people would be most vigilant to the risks for information cascades.

Other contexts may make people more vigilant. For instance, people are mindful of alliance-based biases in testimony. If Jack endorses Jill, his endorsement seems less informative if Jack and Jill are close friends than if they dislike each other; and vice-versa if Jack disparages her. Even children make this inference (Liberman & Shaw, 2020), and by adulthood we use it to reason about the evidential value of consensus: if one of the three eyewitnesses testifying about Jill's alibi as a suspect in a robbery is her friend Jack, his testimony only counts towards consensus if he contradicts her alibi, not if he endorses it (Mercier & Miton, 2019). You may not even expect Jack believe his own testimony; it simply reflects his desire to maintain a relationship with Jill. A similar kind of reasoning may help explain why people so easily dismiss the beliefs of millions of political opponents (Oktar & Lombrozo, 2022): if someone only acquired a belief after hearing it espoused by their party leaders, it's easy to dismiss their belief as dogma.

We suggest that understanding people's reliance on socially-transmitted information requires looking closely at both the epistemic and social motivations *other* people have for accepting or rejecting it. Here, we use the engineering-team scenario introduced earlier to ask people to reason about how much influence early speakers' public "votes" have on subsequent speakers' public votes *and* their private beliefs. Participants are told that initially, four teammates privately believe the blue airplane design is best, and one privately believes the yellow design is best (Figure 1); but, the contest organizers ask them to go around from left-to-right, meaning the yellow voter announces their vote (Yellow) first. Critically, we manipulate participants' perceptions about the speakers' social motivations by introducing the first speaker as "very popular", or omitting mention of their social status.

In Experiment 1, participants first rate which design the second speaker is relatively more likely to vote for after hearing the first speaker's vote, *and* which design they are relatively more confident in privately. We then ask participants to make the same inferences for each subsequent speaker, assuming their previous judgment was correct. For instance, if they expect speaker two to vote yellow, we ask them what speaker three will publicly vote and privately believe after hearing speakers one and two vote yellow. Experiment 2 is similar, but participants are

not shown the speakers' initial beliefs. Instead, we show them that following the first speaker's yellow vote, each subsequent speaker also voted yellow. Participants are then asked which airplane they themselves believed was best.

## Computational Framework

Our computational framework seeks to capture participants' intuitions about the public votes and private beliefs of agents in a sequential voting task. We encode these intuitions into a generative model, and posit that participants can invert this generative model to simultaneously predict agents' public votes and private beliefs throughout the voting sequence.

### Generative Model

Our generative model posits that each agent's vote reflects a mixture of their private beliefs about each option, as well as their desire to gain the social favor of other agents by publicly agreeing with them. The first component- the agent's private belief- is updated throughout the sequence as the agent observes how previous agents have voted. Each agent starts with an initial private belief, represented as a pair of probabilities [$A^0_n$ , $B^0_n$ ], respectively denoting agent n's initial degree of belief that each option A is correct. As agents 1 through n − 1 reveal their public votes, agent n updates their own private belief through a modified form of the social learning rule in Toyokawa et al (2019). Formally, let $V_{-}{}^{A}{}_{n}$ denote the total number of votes for option A among agents 1 through n − 1, and similarly for $V_{-}{}^{B}{}_{n}$. We define agent n's updated belief in option A to be

$$belief_n^A = \frac{w_{self} * A_n^0 + (.1 + V_{-n}^A)^{\theta_d}}{w_{self} * A_n^0 + (.1 + V_{-n}^A)^{\theta_d} + w_{self} * B_n^0 + (.1 + V_{-n}^B)^{\theta_d}} \quad (1)$$

and similarly for B. The parameter $w_{sel\,f} > 0$ denotes the agent's "self-weight," which captures how the agent weighs their own initial belief relative to the public opinions of other agents. If $w_{sel\,f} > 1$, then the agent treats their own initial opinion as if it counts for "more votes" than each other agent's opinion. The "data conformity" parameter $\theta_d$ captures the degree to which the agent's beliefs will conform with a consensus: $\theta_d > 0$ entails a "conformist" agent (i.e.: be more inclined to agree with the option that has the most votes). If $\theta > 1$, then the influence of a consensus on the agent's beliefs will increase superlinearly with the number of agents in the consensus (i.e.: increasing marginal gains).

The second factor that influences an agent's vote is social favor. Intuitively, we treat this term similarly to the belief update rule in equation 1. The key difference is that each previous agent's vote is weighted by that agent's "social power" $p_k$. For example, if agent k is especially popular or influential, then k will have a higher social power value than the other agents. Formally, let $P_{-}{}^{A}{}_{n} = \sum_{k<n} I[Vote_k = A] * p_k$, i.e.: the weighted sum of votes for agents 1 through n − 1, weighted by the social power of each agent. We define the social influence of votes 1 through n − 1 on agent n as: and similarly for $influence^{B}_{n}$. The parameter $\theta_s$ is the agent's "social conformity," which issimilar to the "data conformity" parameter $\theta_D$, but captures the degree to which the agent is influenced by the social power associated with previous votes.

$$influence_n^A = \frac{(.1 + P_{-n}^A)^{\theta_s}}{(.1 + P_{-n}^A)^{\theta_s} + (.1 + P_{-n}^B)^{\theta_s}} \quad (2)$$

Given these two update rules, we assume that the probability that agent n will vote for option A is a weighted sum of agent n's private belief in A ($belief_n^A$) and the social influence for option A on agent n ($influence^{A}_{n}$), i.e.:

$$P(Vote_n = A) = w_{acc} * belief_n^A + (1 - w_{acc}) * influence_n^A \quad (3)$$

The weight parameter $w_{acc}$ captures the degree to which agent n's vote is motivated by accuracy (i.e.: reflects their private belief in option A) versus social favor (i.e.: reflects a desire to align their public opinion with the set of agents voting for A). Each agent is thus defined by the five parameters $w_{sel\,f}$ , $\theta_d$ , $\theta_s$ , $w_{acc}$ , and p. For the purpose of our initial experiments, we assume that all agents have equal values for the first four parameters, but that social power may vary between agents.

### Inference and Prediction

Given each agent's parameter values for each agent (denoted Θ) and initial belief (denoted β), we can use equations (1)-(3) to probabilistically generate a sequence of votes and beliefs by iterating the model forward: In step k, given votes $V_1,\ldots,V_{k-1}$, we compute agent k's updated beliefs, as well as the social influence on k from previous votes, then sample agent k's vote according to equation (3). This defines a probability distribution over vote sequences $\bar{V}_N$ and belief sequences $\bar{B}_N$ , i.e.: $P(\bar{V}_N , \bar{B}_N |\Theta, \beta)$. In Experiment 1, participants produce this exact sequence of judgments: they are first shown each agent's initial belief in each option, then they sequentially predict each agent's degree of belief in each option, and the predict that agent's vote. We then implement a Bayesian parameter estimation procedure (see next section for details) to infer the set of parameter values Θ that best explain each participant's judgments.

### Parameter Estimation

In Experiment 1, each participant produces a sequence of vote probabilities $v_2,\ldots,v_5$ and degrees of belief $b_2,\ldots,b_5$ in each option. To fit the model parameters to participant data, we implemented a Hierarchical Bayesian estimation procedure which simultaneously estimates parameter values $\Theta_j$ for each individual participant, as well as a group-level distribution $P(\Theta|\alpha)$ over parameter values. This group-level distribution is defined by hyper-parameter α, which is drawn from a prior distribution P(α). We can then compute the joint posterior distribution over $\bar{\Theta}$ (the set of parameter values for each participant) and α (the global hyper-parameters) as

$$P(\bar{\Theta}, \alpha|D) \propto P(D|\bar{\Theta}, \alpha)P(\bar{\Theta}|\alpha)P(\alpha) \quad (4)$$

We estimated this distribution via an MCMC sampling procedure iterated for 100,000 samples. Convergence was validated using a Gelman-Rubin statistic computed with 4 parallel chains, using a threshold of 1.1. We then took MaP estimates of $\Theta_j$ for each participant, as well as MaP estimates of $\alpha$ for each parameter.

## Experiment 1

In Experiment 1, we presented mTurkers with the scenario above (color-coding the airplanes yellow or blue, to prevent participants from seeing the designs for themselves). We manipulated social favor by either describing the first speaker as "*very popular*" or omitting mention of status.

**Participants**. We recruited n=118 mTurkers in one of two conditions (Popular or Anonymous). An additional 10 were screened out prior to participating for failing basic comprehension checks about the instructions; an additional 2 were excluded after participating for failing an attention check and providing gibberish explanations.

**Procedure**. Participants were introduced to a five-person team building a remote control airplane for a contest using one of two designs. The designs were presented in color-coded (yellow or blue) boxes in order to prevent participants from evaluating them directly. They were told that the contest organizers asked each person to look in each box and think silently about which would be best for the contest (though "looking in the box" is probably unnecessary to ensure adults infer private knowledge, it makes the procedure more consistent with a procedure being used with children). A thought-bubble next to each person showed their private belief: one person privately believed the yellow plane was best, while the remaining four all believed the blue plane was best. Next, participants were told that the contest organizers asked the teammates to go around one-by-one from left to right and say which kind of airplane they thought was best. Then, they were told that (given the left-to-right order), the first person to speak was the teammate who privately believed yellow, and they voted yellow. In the Anonymous condition, the teammate was left unnamed; but in the Popular condition, he was introduced as "Max", and described as being very popular.

Participants were then asked to infer (1) what the second speaker would say, after hearing Max say yellow, and (2) what the second speaker privately believed. On each subsequent trial, they were asked to assume their answers for the previous round were correct. For instance, if they answered that the second speaker would also say yellow, they were asked (1) what the third speaker would say after hearing Max and the second speaker say yellow, and what (2) what the third speaker would privately believe. Participants responded using a 20-point scale for all questions anchored at 1-"definitely blue" and 20-"definitely yellow", and were asked to briefly explain their answer to the final decision question.

**Results and Discussion.** Participants provided ratings for each of four voters (as the first vote was specified by the stimulus), producing a probability for each vote $v_2,...,v_5$ and corresponding degree of belief for each vote $b_2,\ldots,b_5$. Our analysis of these data focused on two primary questions. First, do participants expect later voters to be influenced by a desire to gain the social favor of previous voters, in addition to the voter's own private beliefs? In the absence of this social influence, participants should expect each voter's private beliefs to align with their public vote. That is, participants who do not expect social favor to matter should respond that the likelihood a voter will vote for "yellow" is exactly equal to the voter's degree of belief that yellow is the correct answer. Thus, we evaluated systematic differences between participant judgments about public votes and private beliefs.

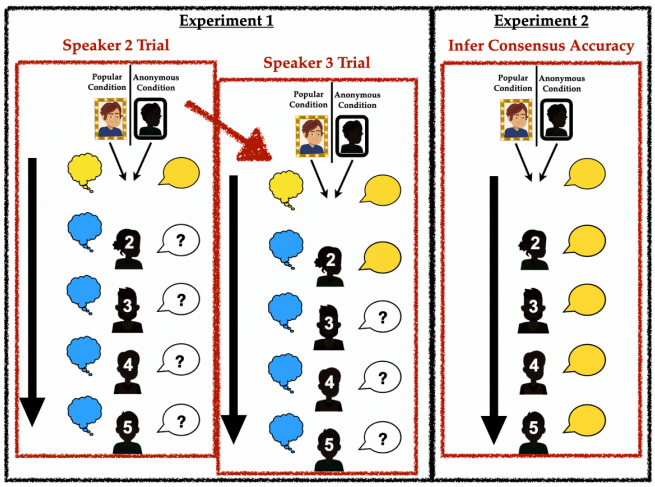


**Figure 1**: Procedure and example stimuli for Exps 1 & 2. In Exp 1, thought bubbles depict each agent's private beliefs, while speech bubbles depict public votes. Participants see all agents' private beliefs & the first agent's public vote. They then predict each subsequent agent's public vote and private belief, iterating through remaining agents. In Exp 2, participants only see agents' public votes and rate their own trust in the public consensus.

As shown in Figure 2a, mean participant responses revealed a significant gap between vote probability ratings and degree of confidence ratings: in particular, participants rated each voter as less likely to privately believe yellow (agreeing with the first voter) than vote yellow publicly ($\beta_{Think}$ = -2.50, SE = .36, $p$<.001). This trend is consistent across all 4 speakers and across both conditions. Furthermore, this gap was substantially and significantly larger in the "popular" condition than the "anonymous" condition ($\beta_{Popular}$ = 5.06, SE = .50, $p$<.001). That is, when the first speaker is described as having a disproportionate amount of social power, participants reported that the subsequent voters were substantially more likely to vote in line with the first voter, but were not more substantially likely to believe that the first voter is correct. Thus, participants expect the popular voter to have more direct influence over the public votes of other agents, but not the private votes of those agents. This is consistent with our

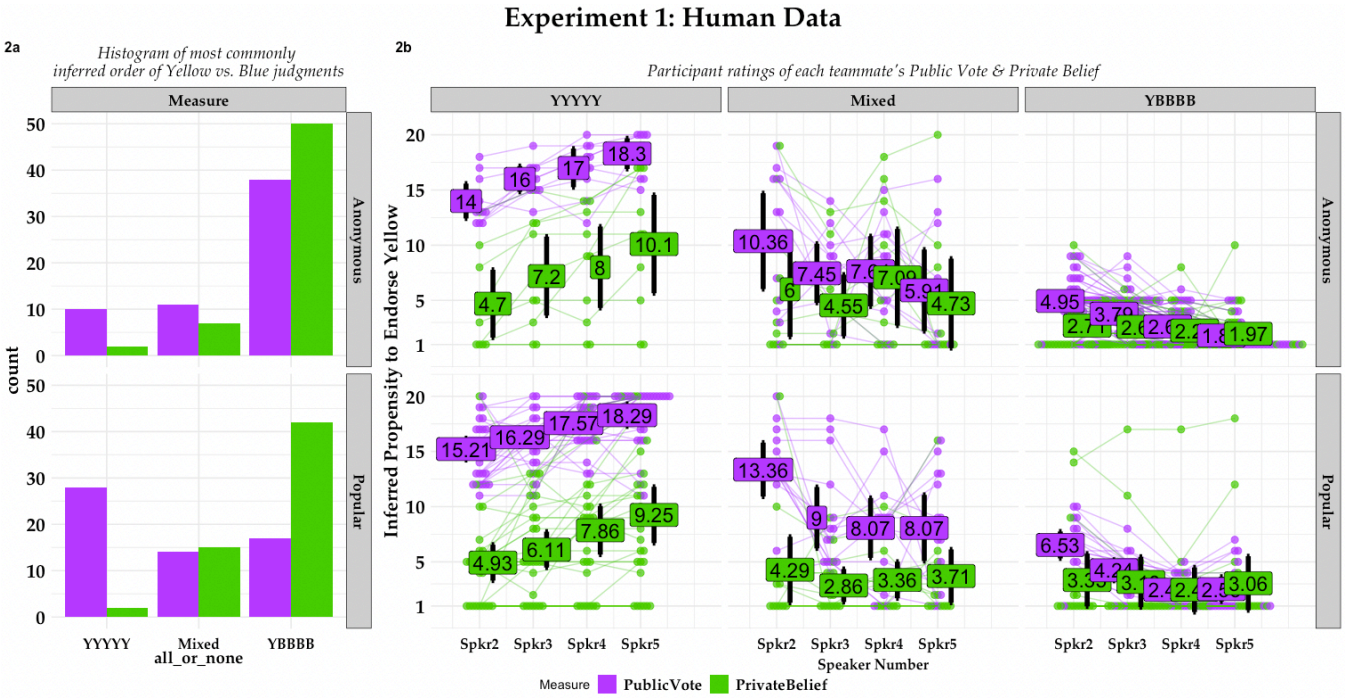


**Figure 2**: Experiment 1 results. Color denotes inferred PublicVote or PrivateBelief. **Panel 2a:** Histogram of predicted vote and belief sequences. For visualization, participants are groups by the vote sequence they predict: YYYYY ("all speakers will vote (or believe) yellow"), YBBBB ("only Spkr_1 will vote (or believe) yellow"), and "Mixed" (all other sequences). **Panel 2b**: lines join each participant confidence in inferring each speaker's public vote and private belief; values denote means.

hypothesis that participants expect votes to reflect both private beliefs and a desire for social favor.

A second focus of our analysis was the presence of information cascades, wherein later voters override their initial private beliefs and vote in line with previous voters. As shown in Figure 2b, average participant responses do reflect an expectation of information cascades. In particular, participants who predicted that the second voter would "flip" (i.e.: vote contrary to their initial private belief, and instead vote in line with the first voter) also consistently predicted that subsequent voters would also "flip," with increasing probability. This effect was consistent across both the anonymous and popular conditions. However, a significantly larger proportion of participants predicted this flip in the *Popular* condition than the *Anonymous* condition. Thus, not only did participants anticipate information cascades, they were more likely to predict a cascade when the first voter had a disproportionate amount of social power.

To better understand the intuitions underlying participant judgments, we estimated optimal parameters for each individual participant. Model predictions fit participant data well (R=.85 for "anonymous" condition, R=.87 for "popular" condition). This enabled us to analyze the distribution of optimal parameters for explaining participant responses (see Figure 3). Since participants who predicted speaker two would flip were overwhelmingly likely to predict all subsequent voters would flip as well, we plotted separate distributions for the two groups of participants. Across both conditions, we found that participants who predicted a flip had significantly lower estimated values for $w_{self}$, which corresponds to the expectation that votes are determined primarily by a desire for social favor, rather than epistemic accuracy. These participants also had significantly higher estimates for the social power of the first voter, explaining the disproportionate amount of influence these participants attributed to the first voter. Additionally, participants in the "popular" condition had significantly higher estimates for the first speaker's social power than in the "anonymous" condition.

## Experiments 2a-b

Since participants' explanations in a pilot suggested that some participants might believe that even an Anonymous first speaker's answer would influence subsequent speakers, we ran two versions of Experiment 2. In Experiment 2b, we told participants that the contest organizers instructed the teammates to write down which design they believed was best on a piece of paper after thinking silently; in Experiment 2a, we omitted mention of this, leaving the contest organizer's instructions identical to Experiment 1. We expected this would make participants in both conditions more likely to infer that teammates would stick to their initial beliefs when asked to vote publicly, but that participants in the Popular condition would nevertheless be more skeptical of the consensus than participants in the Anonymous condition.

**Participants**. Experiment 2a comprised n=81 MTurkers in one of two conditions (Popular or Anonymous); an

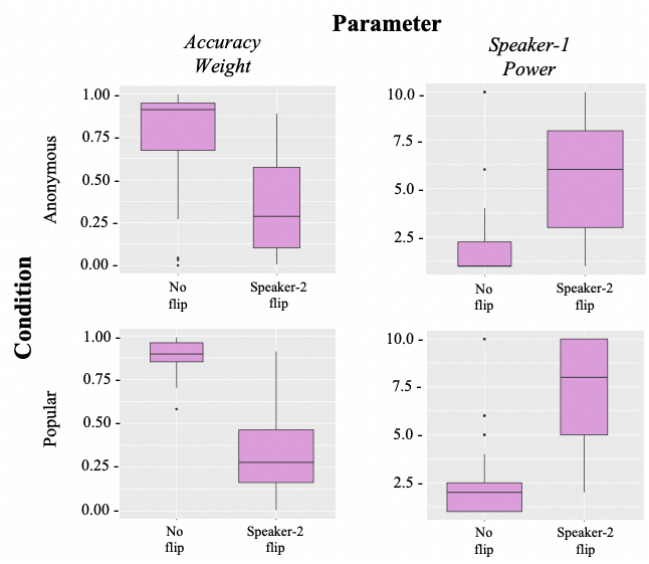


**Figure 3**: Optimal parameters from individual model-fitting. **Left column**: $w_{acc}$ parameter (relative weight of private belief vs desire for social favor on each agent's vote). **Right column**: first speaker's social power. **Rows**: Anonymous vs. Popular condition fits. **X-axes**: participants split according to their "Spkr2_Flip" prediction: would speaker 2 vote contrary to their initial private beliefs or not?

additional 14 were screened out prior to participating for failing basic comprehension checks about the instructions. Experiment 2b comprised n=80 in one of two conditions (Popular or Anonymous); an additional 16 were screened out for the same reason.

**Procedure**. As in Experiment 1, participants were introduced to a five-person team building a remote control airplane for a contest using one of two designs. However, participants in Experiment 2 were not shown which design each participant initially believed best. And, after introducing the first speaker as "Max" and describing him as "very popular" (Popular condition) or omitting mention of his status (Anonymous condition), participants saw that every speaker publicly voted yellow, one-by-one. They then rated which airplane they themselves believed was best, and predicted which design the team would choose, briefly explaining their answers for each. Participants responded using a 20-point scale anchored at 1-"definitely blue" and 20-"definitely yellow".

**Results and Discussion.** In order to compare trust in the Anonymous condition to indifference between the two designs, we centered the response scale at the midpoint of the 20-point scale. When the first speaker's status was omitted (Anonymous), participants believed the consensus-endorsed yellow design was best (*Anonymous*: $\beta_{Intercept}$ = 5.15, SE = .52, $p$<.001). However, when the first speaker was described as popular, participants were significantly less confident in the yellow design, despite the team's unanimous endorsement (*Popular*: $\beta_{Popular}$ = -2.53, SE = .72, $p$<.001), though they were more confident in the yellow design than the blue design (*Popular*: $\beta_{FlipIntercept}$ = 2.62, SE = .50, $p$<.001). Results were similar when we compared Experiment 2a and 2b directly, with no effect of experiment version ($\beta_{Anonymous}$ = 4.91, SE = .74, $p$<.001; $\beta_{Popular}$ = -2.48, SE = 1.02, $p$<.017; $\beta_{ExpNum}$ = 0.49, SE = 1.04, $p$=.64; $\beta_{PopularExpNum}$ = -0.10, SE = 1.45, $p$=.95). Participants were also confident that the team would decide on the yellow design after talking together, with no difference between conditions (*InferBest*: $\beta_{Anonymous}$ = 7.04, SE = .35, $p$<.001; $\beta_{Popular}$ = 0.27, SE = .48, $p$=.58).

## General Discussion

Briefly put, our results suggest that people make systematic inferences about how social status and consensus affect their informants' public and private judgments, and even recognize the role of early speakers' votes in triggering information cascades (Banerjee, 1992). Notably, out of the 118 participants in Exp 1, *none* of the 62 who believed that Spkr 2 would stick to their original beliefs (i.e., vote blue) expected any subsequent speakers to flip their vote; but of the 56 participants who believed Spkr2 *would* flip (40 in Popular and 16 in Anonymous), 48 expected at least one more to flip as well. In other words, only participants who expected Spkr1 to flip Spkr2 expected consensus to flip, and few expected Spkr2 to flip unless Spkr1 was Popular. This belief in the teammates' tendency to stick to their beliefs unless pressured by consensus or high-status individuals may help explain why participants in Exp 2 trusted the Anonymous consensus despite the risk of social influence. It may also help explain why people sometimes appear credulous of 'false' consensus (Yousif, et al., 2019), but are more discerning when dependencies between informants are emphasized (Desai et al., 2022) or when informants appear to be conscientiously endorsing a belief instead of parroting it or conforming to social pressure (Alister et al., 2022; Richardson & Keil, 2022a, 2022b; Mercier & Miton, 2019), and even anti-consensus after falling prey to conspiracy theories (Light et al., 2022; cf. Oktar & Lombrozo, 2022). How so?

Since people's informants are no more capable of evaluating every claim on the merits than they themselves are, almost any real-world consensus will be downstream of *some* dependencies. And these dependencies can make consensus more accurate instead of less. So instead of disregarding consensus simply because informants *might* have influenced each other, people may often count on their informants to be discerning judges of each others' testimony unless given reason to believe otherwise. In other words, even if people are less skeptical of dependencies in ball-and-urn tasks than normative models prescribe (Whalen et al., 2018; Xie & Hayes, 2022), or overly credulous of 'false' consensus in news reports (Yousif, et al., 2019), it may be because such tasks don't capture the kinds of dependencies that people's information environments have made them vigilant towards. Our results suggest that reasoning about the effects prestige and conformity have on their informants' reliability may give people reason to be more skeptical.

## References

Altay, S., Claidière, N., & Mercier, H. (2020). It happened to a friend of a friend: Inaccurate source reporting in rumour diffusion. *Evolutionary Human Sciences*, *2*, e49. https://doi.org/10.1017/ehs.2020.53

Anderson, L. R., & Holt, C. A. (1997). Information Cascades in the Laboratory. *The American Economic Review*, *87*(5), 17. https://www.jstor.org/stable/2951328

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). https://doi.org/10.1038/s41562-017-0064

Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, *107*(3), 797–817. https://doi.org/10.2307/2118364

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. https://doi.org/10.1086/261849

Boyd, R., & Richerson, P. J. (1995). Why does culture increase human adaptability. *Ethology & Sociobiology*, *16*, 125–143. https://doi.org/10.1016/0162-3095(94)00073-G

Desai, S. C., Xie, B., & Hayes, B. K. (2022). Getting to the source of the illusion of consensus. *Cognition*, *223*, 105023. https://doi.org/10.1016/j.cognition.2022.105023

Dietrich, F., & Spiekermann, K. (2013). Epistemic Democracy with Defensible Premises. *Economics and Philosophy*, *29*, 34. https://doi.org/doi:10.1017/S0266267113000096

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive Theories. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199399550.013.28

Jiménez, Á. V., & Mesoudi, A. (2019). Prestige-biased social learning: Current evidence and outstanding questions. *Palgrave Communications*, *5*(1), 20. https://doi.org/10.1057/s41599-019-0228-7

Kameda, T., Toyokawa, W., & Tindale, R. S. (2022). Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-022-00054-y

Kao, A. B., Miller, N., Torney, C., Hartnett, A., & Couzin, I. D. (2014). Collective Learning and Optimal Consensus Decisions in Social Animal Groups. *PLoS Computational Biology*, *10*(8), e1003762. https://doi.org/10.1371/journal.pcbi.1003762

Laan, A., Madirolas, G., & de Polavieja, G. G. (2017). Rescuing Collective Wisdom when the Average Group Opinion Is Wrong. *Frontiers in Robotics and AI*, *4*. https://doi.org/10.3389/frobt.2017.00056

Liberman, Z., & Shaw, A. (2020). Even his friend said he's bad: Children think personal alliances bias gossip. *Cognition*, *204*, 104376. https://doi.org/10.1016/j.cognition.2020.104376

Mannes, A. E. (2009). Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision. *Management Science*, *55*(8), 1267–1279. https://doi.org/10.1287/mnsc.1090.1031

Mercier, H., & Miton, H. (2019). Utilizing simple cues to informational dependency. *Evolution and Human Behavior*, *40*(3), 301–314. https://doi.org/10.1016/j.evolhumbehav.2019.01.001

Morgan, T. J. H., Rendell, L. E., Ehn, M., Hoppitt, W., & Laland, K. N. (2012). The evolutionary basis of human social learning. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1729), 653–662. https://doi.org/10.1098/rspb.2011.1172

Oktar, K., & Lombrozo, T. (2022). Mechanisms of Belief Persistence in the Face of Societal Disagreement. *Proceedings of the Cognitive Science Society*, 8.

Pilditch, T. D., Hahn, U., Fenton, N., & Lagnado, D. (2020). Dependencies in evidential reports: The case for informational advantages. *Cognition*, *204*, 104343. https://doi.org/10.1016/j.cognition.2020.104343

Pink, S. L., Chu, J., Druckman, J. N., Rand, D. G., & Willer, R. (2021). Elite party cues increase vaccination intentions among Republicans. *Proceedings of the National Academy of Sciences*, *118*(32), e2106559118. https://doi.org/10.1073/pnas.2106559118

Raafat, R. M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences*, *13*(10), 420–428. https://doi.org/10.1016/j.tics.2009.08.002

Richardson, E., & Keil, F. C. (2022a). Anger, evidence, & trending opinions: We trust consensus when we believe it reflects genuine persuasion. PsyArXiv. https://doi.org/10.31234/osf.io/8gkqa

Richardson, E., & Keil, F. C. (2022b). The potential for effective reasoning guides children's preference for small group discussion over crowdsourcing. *Scientific Reports*, *12*(1), 1193. https://doi.org/10.1038/s41598-021-04680-z

Toyokawa, W., Whalen, A., & Laland, K. N. (2019). Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, *3*(2), 183–193. https://doi.org/10.1038/s41562-018-0518-x

Whalen, A., Griffiths, T. L., & Buchsbaum, D. (2018). Sensitivity to Shared Information in Social Learning.

*Cognitive Science*, *42*(1), 168–187. https://doi.org/10.1111/cogs.12485

Xie, B., & Hayes, B. (2022). Sensitivity to Evidential Dependencies in Judgments Under Uncertainty. *Cognitive Science*, *46*(5). https://doi.org/10.1111/cogs.13144

Yousif, S. R., Aboody, R., & Keil, F. C. (2019). The Illusion of Consensus: A Failure to Distinguish Between True and False Consensus. *Psychological Science*, *30*(8), 1195–1204. https://doi.org/10.1177/0956797619856844