Author(s)
Emory Richardson
Isaac Davis
Frank Keil

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?

The question of interest is whether reasoning about how social power and conformist tendencies affect individuals' *public* and *private* judgments would lead people to predict information cascades and pluralistic ignorance in a sequential voting scenario. More concretely: suppose that 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 publicly. Participants are told that initially, four teammates privately believe the blue airplane design is best, and one privately believes the yellow design is best; but, the contest organizers ask the teammates one-by-one from left-to-right, and participants are shown that the first speaker's public vote is the same as their initial belief.

Our critical manipulation is whether we describe the yellow "dissenter" as being "very popular" with their four teammates or omit mention of their social status, and whether the dissenter votes first (Exp 1) or last (Exp 3). This creates four conditions. Our primary prediction is that when the dissenting speaker ***both*** (A) speaks first ***and*** (B) is popular, participants will rate subsequent speakers as more likely (relative to participants' ratings in the other three conditions) to:

(1) not only shift their private beliefs toward yellow (i.e., grow less confident in blue), but
(2) "flip" their public votes to yellow even if they still privately believe blue is better, thus
(3) creating a greater split between PublicVotes & PrivateBeliefs in the Popular+SpeaksFirst condition than in the other three conditions.

We outline these predictions more precisely using a computational model described in Section 5 below.
In short, we predict that the "social model" described in Section 5 will better fit participants' ratings in the "Dissenter_isPopular+SpeaksFirst" condition, while the "asocial model" and social model will fit participants' ratings equally well in the other 3 conditions.

Because the final degree of consensus in Experiments 1 and 3 depends on participants' inferences, Experiment 2 stipulates a unanimous consensus (see below for details). We predict that participants will trust the design endorsed by the unanimous consensus in both the FirstSpeaker_isPopular and FirstSpeaker_isAnonymous condition, but critically, trust will be significantly attenuated when the first speaker is popular compared to when they're anonymous.

3) Describe the key dependent variable(s) specifying how they will be measured.

After seeing all 5 teammates' initial Private Beliefs and Speaker1's PublicVote, each participant in Experiments 1 and 3 will rate the likelihood that Speaker2 will (A) Publicly Vote and (B) Privately Believe blue or yellow, given Speaker1's PublicVote. They'll then be asked to rate the PublicVote and PrivateBelief of each subsequent speaker, while assuming that their inferences about the previous speakers were correct.

Following these 8 ratings, participants will be told that after each teammate voted, the team was asked to talk together to make a final decision about the design; participants will be asked to infer (1) the team's final decision and (2) which design they themselves believe is best.

In Experiment 2, participants will not see any teammates' PrivateBeliefs, but will see all 5 teammates Publicly Voted yellow, one-by-one, producing a unanimous Public Vote. As in Experiments 1 and 3, they'll then be told that after each teammate voted, the team was asked to talk together to make a final decision about the design; participants will be asked to infer (1) the team's final decision and (2) which design they themselves believe is best.

All ratings will be made on a 20-point scale.

Participants will be asked to briefly explain their ratings of the the team's final decision and which design they themselves infer is best.

4) How many and which conditions will participants be assigned to?

Participants in Experiment 1 will be assigned to one of two conditions:

Dissenter_isPopular+SpeaksFirst
Dissenter_isAnonymous+SpeaksFirst

Participants in Experiment 2 will be assigned to one of two conditions:

UnanimousVote_isPopular+SpeaksFirst
UnanimousVote_isAnonymous+SpeaksFirst

Participants in Experiment 3 will be assigned to one of two conditions:

Dissenter_isPopular+SpeaksLast
Dissenter_isAnonymous+SpeaksLast

5) Specify exactly which analyses you will conduct to examine the main question/ hypothesis.

We are pre-registering multiple analyses. The analysis based on our computational model for Experiments 1 & 3 is planned as follows.

Each participant in each condition will produce 8 numerical responses: a prediction about the likelihood that each of four agents will vote for yellow or blue, and a prediction about the strength of each agent's private belief that yellow or blue is the correct answer. As part of our analysis, we use two different computational models to generate predictions for each of these same values. These models simulate a sequential voting process that iterates through the following steps:

1. Agent n observes the public votes of agents 1 through n-1, and updates their own private belief based on these votes, as well as agent n's initial belief (which is fixed as part of the experimental condition). This belief updated is computed according to the following equation:

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

   Here, A_n^0 is agent n's initial degree of belief in option A, and similarly for B_n^0. The parameter w_{self} captures how much the agent weighs their own initial opinion relative to the public opinions of other agents. θ_d is a *data conformity* parameter, which captures how responsive agent n is to a public consensus: higher values of θ_d indicate a stronger conformity to consensus. V_{-n}^A counts the number of previous votes for A, and similarly for V_{-n}^B.

2. Agent n then computes the *social influence* associated with each option based on currently public votes, according to the following equation:

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

   Here, θ_s is a *social conformity* parameter, which captures how responsive the agent is to social influence. P_{-n}^A captures the *social power* associated with currently public votes for option A (and similar for P_{-n}^B). Each agent has a social power value, which affects how much social influence their public vote carries. In the *asocial* model (Model 1), each agent's social power is fixed to 1. In the *social power* model (Model 2), the named agent's social power value is treated as a free parameter and fit to data.

3. Agent n then computes their overall probability of voting for option A as a weighted sum of equations (1) and (2):

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

   Here, w_{acc} is a free parameter between 0 and 1 which captures how much the agent values accuracy (i.e.: voting in line with their own private beliefs) versus social favor

(i.e.: voting in line with the most socially powerful group): w_{acc}=0 denotes an agent who only cares about social influence, while w_{acc}=1 denotes an agent who only cares about accuracy. After computing this probability, the agent then votes for option A with probability P(Vote$_n$=A), or B with probability 1-P(Vote$_n$=A).

The asocial model thus has four free parameters (w_{self}, θ_d, θ_s, and w_{acc}), while the social model has an additional social power parameter P. After collecting participant data, we optimize each model to each participant individually by minimizing the mean squared error between model predictions and each participant's responses, separately for each model. We hypothesize that the "social" model (where the named agent's social power is treated as a free parameter) will outperform the "asocial" model (where all agents have fixed and equal social power) in only one of our four conditions: popular x first (i.e.: the named agent is described as popular and is the first voter). We evaluate this hypothesis in two ways: first, for each condition, we apply a t-test to determine whether there is a significant difference between the distributions of mean-squared errors computed using each model. Second, we compute the aggregate correlation between each model's predictions and participant responses for each condition, then compute a bootstrapped 95% confidence interval over the difference in model correlations. We interpret a 95% CI entirely above 0 to indicate a positive significant difference in model correlation, and a 95% CI containing 0 to indicate no significant difference in model correlations.

All models and optimizations were implemented using WebPPL, a probabilistic programming language for generative models. The attached zip file includes:

- A text file Agenda_setting_paramFitting.txt with the code for our models and optimization algorithm.
- Two csv files, exp1mod_pop_data.csv and exp1mod_anon_data.csv, containing participant response data from a previous pilot study, included as a working example of how our model and optimization algorithm work.
- An R script Exp1mod_analysis.R that generates the plots and implements the significance testing and bootstrapping described above.

In addition to the computational model described above, we will use the following model of participants' ratings of each teammate's PublicVote and PrivateBelief to test whether participants believe that teammates in the Popular condition are less likely to believe privately yellow than vote yellow publicly:

mod_SayThinkDifference: ratings ~ judgmentType_SayThink*PowerLevel_PopAnon + (1| subID:SpkrID)

mod_TrustMaj: In Experiments 1 and 3, we will analyze participants' confidence in the majority judgment using model comparison. We predict that model A below will outperform any combination of the three factor models in (A), (B), (C), where the "inferAvg_" predictors are the average of the 4 ratings for Speakers 2-5 (for PrivateBelief or PublicVote, respectively).

(A) trustMaj ~ inferAvg_PrivateBeliefs
(B) trustMaj ~ inferAvg_PublicVotes
(C) trustMaj ~ PowerLevel_PopAnon

In Experiment 2, where PublicVotes are stipulated and PrivateBeliefs are not revealed to participants, we will test whether participant are less trusting of unanimous consensus that emerges following a Popular first speaker than an Anonymous first speaker using the following modification of model (C) above. By subtracting the midpoint of the scale (10.5) from the 20-point ratings, the intercept of the model is equivalent to a one-sample t.test participants' trust or distrust in the unanimous consensus, while the PowerLevel_PopAnon reveals whether these ratings differ by condition.

mod_distrustUnanimous: (trustMaj — 10.5) ~ PowerLevel_PopAnon

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

Participants will be required to pass an attention check after reading the initial instructions, consisting of basic comprehension questions about the instructions. Participants who fail the attention check twice will be screened out of the study. In order to counter "bot-farmer" responders to online surveys, participants explanations of (A) which design they believe the team will choose and (B) which design they themselves believe is best) will be separated from the data and hand-coded as "suspicious" or "clean". Given prompts (A) and (B) above, examples of "suspicious" vs. "clean" responses (taken from pilots) are:

Suspicious: "*I made my own decision*", *"This one is perfect", "That are a best situations",* or *"I just feel nothing too much and i am guessed that"*
Clean: *"It will be hard for the lone dissenter to change the others minds after they've learned that they all agree with each other", "The majority chose blue, so it will be easier to convince the other person",* or *"Everyone on the team idolizes Max and wants to be in his good standing. Since Max thinks the best plane is yellow, a good chunk of the team is likely going to change their choice to yellow to reflect Max."*

Analyses will be run both with and without these responses and compared (see secondary analyses below).

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about <u>exactly</u> how the number will be determined.

Sample sizes were decided based on effect sizes in piloting. Our target sample size is n=100 participants per condition in each Experiment, after exclusions; we will recruit n=120 per condition to account for exclusions based on exclusion rates in previous pilots.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

When participants are asked for qualitative explanations of their judgments, reviewing these explanations by hand makes it fairly obvious that a subset are generated by chatbots or bad-faith "survey farmers" that are difficult to screen out by advance criteria. As a secondary analysis, we will compare the model fit when these responses are included with the "clean" data, as described in Section 6.