How Social Influence Shapes Collective Intelligence in Binary Choices: Reconciling Experimental Disparities with Model

Vicky Chuqiao Yang[1, 2] and Levi Grenier[1]

[1]MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA

[2]MIT Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA

Abstract

Understanding when groups make good or bad collective decisions is a central societal concern, with social influence being a key factor. Experimental findings on its effects are mixed—some suggest social influence helps, others that it hurts, and some suggest it depends on a myriad of factors. We reconcile these disparate conclusions for binary choice tasks by proposing a simple mathematical model that captures individuals integrating independent judgment and social information under various experimental structures. Our model predicts a bifurcation, the emergence of two possible outcomes for group composition. By analyzing data from four published experiments, we demonstrate that the predicted bifurcation has been present in some prior experiments. The model also reproduces several disparate experimental findings: (1) The accuracy of collective decisions is nonlinearly influenced by the accuracy of initial judgments; (2) Social influence can enhance individual accuracy while reducing collective accuracy; (3) Groups can exhibit self-correcting dynamics, avoiding lock-in of inferior options. Using the model, we predict that some of these effects hold only under specific conditions, and identify parameters where we expect them to change. We then use our model to derive parameter regions under which we expect social influence to improve or hinder collective accuracy. Notably, while the psychological mechanisms remain consistent, the experimental structure—sequential or synchronous updating and task difficulty—is critical in shaping outcomes. Our findings suggest that disjointed and seemingly contradictory results can be explained through simple, first-principle models involving nonlinear interactions, offering a potential solution to reproducibility challenges in collective

intelligence research.

1 Introduction

Human groups are often capable of outperforming their individual members (Surowiecki, 2005), yet they have also collectively made decisions widely regarded as poor, such as companies making erroneous strategic decisions and nations enacting disastrous political policies. With the identification of a collective intelligence factor—where some groups consistently perform better across a variety of tasks (Woolley et al., 2010), researchers have extensively investigated the conditions that either enable or hinder group performance. A critical factor identified is social influence, whereby individuals' beliefs or behaviors are affected by those of others. Despite extensive studies, the literature presents disparate and sometimes conflicting findings on how social influence affects collective intelligence.

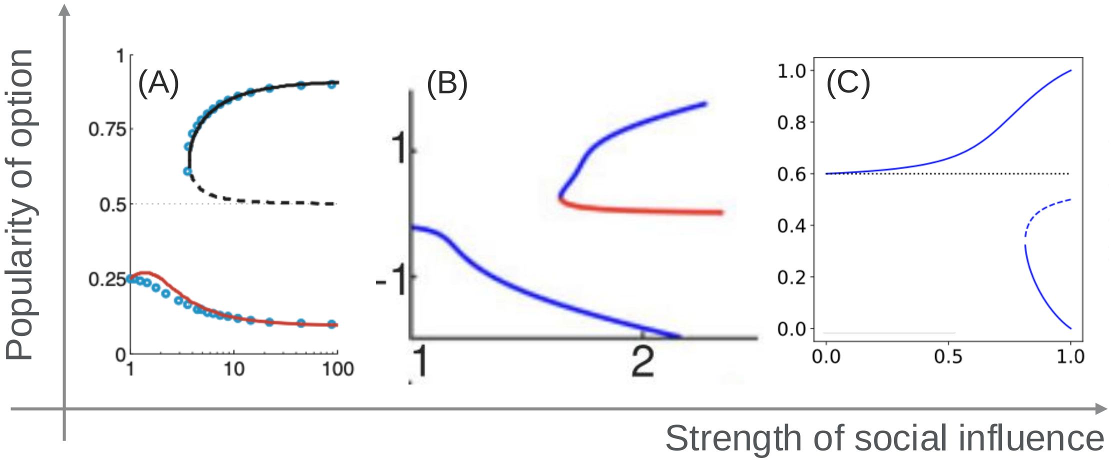
Some experimental studies conclude that social influence negatively impacts collective intelligence (Da & Huang, 2020; Frey & van de Rijt, 2021; Lorenz et al., 2011), and groups would achieve greater accuracy if individuals arrive at their judgments independently, which are then aggregated by taking the mean, median, or majority vote. The underlying mechanism posited is that independent errors tend to be uncorrelated and thus cancel out during aggregation. Social influence can lead to herding behavior, reducing the diversity of opinions and consequently increasing error correction (Becker et al., 2017; Hong & Page, 2004). However, other experimental studies have found the opposite effect—social influence aids individuals in refining their personal judgments, and when these are aggregated, it enhances the accuracy of collective conclusions (Farrell, 2011; Gürçay et al., 2015; Jayles et al., 2017).

Given these conflicting conclusions, numerous studies have argued that, as with many questions in the social and behavioral sciences, it depends. A wide range of factors have been identified as shaping these outcomes, including the structure of interaction networks and their adaptability (Almaatouq et al., 2020; Becker et al., 2017), the distribution of initial judgments (Almaatouq et al., 2022; Becker et al., 2022; Frey & van de Rijt, 2021), whether individuals report independent judgments prior to interaction (Minson et al., 2018), aggregation method (Kao et al., 2018), and evaluation metrics (Frey & van de Rijt, 2021).

Even for the narrowly defined task of binary choice over factual questions, studies have found disparate conclusions. For instance, one study found social influence increases individual accuracy—defined as the likelihood of any individual choosing the correct answer——but decreases collective accuracy, or the likelihood of the group majority choosing the correct answer (Frey & van de Rijt, 2021). Another study revealed the effect of social influence depends on the group's initial accuracy: when initial accuracy exceeds 50%, social influence tends to enhance it, but when initial accuracy is below 50%, social influence has a detrimental effect (Becker et al., 2022). Additionally, groups can exhibit self-correcting behavior—when an inferior option gains popularity, the group will choose it with a probability lower than its popularity, thereby reducing its appeal and avoiding lock-in of the inferior option (van de Rijt, 2022). However, this self-correcting effect is not guaranteed in human group dynamics and inferior options could persist in difficult tasks (Frey & van de Rijt, 2021). While these studies offer valuable insights, the lack of a cohesive theoretical framework has become a bottleneck to generating broadly applicable, actionable insights.

While the experimental findings remain disparate, theoretical models investigating the role of social influence in groups' binary choice outcomes across various disciplines and methods have strikingly converged on a common conclusion. These models consider groups making binary choices under social influence across diverse contexts, encompassing both human and non-human animal groups, including human economic decisions (Brock & Durlauf, 2001), human opinion formation (Mori et al., 2012; Yang et al., 2021), bees selecting nesting sites (Gray et al., 2018), and fish choosing paths to locate food (Couzin et al., 2011). The modeling methods span agent-based simulations, utility maximization, and differential equations, each incorporating distinct contextual assumptions. For instance, some models assume interactions are mediated by spatial proximity, while others do not. Despite variations, these theoretical approaches all predict that as social influence surpasses a critical threshold, a bifurcation occurs, indicating that the number of possible group compositions increases from one to two. With low social influence, the models predict a single group composition—determining the proportion of the group choosing an option over the alternative. Conversely, with high social influence, the models predict two possible outcomes, and which outcome occurs is subject to the interaction of the initial conditions with the basins of attraction of the two equilibria. Figure 1 compares three figures, each extracted from a theoretical modeling study, to demonstrate this striking convergence across the literature. A summary of all five studies reviewed is shown in Table 1.

Figure 1: Figures from three prior modeling studies illustrate a remarkable convergence in the theoretical literature on the effect of social influence on collective outcomes for binary choices.



Notes. (A) From Couzin et al. (2011), for fish selecting travel directions to locate food. (B) From Gray et al. (2018), for honey bees choosing nesting sites. (C) From Yang et al. (2021) for human opinion formation. Despite differences in variable names across studies, the horizontal axes represent the strength of social influence, while the vertical axes indicate the popularity of one option within the group. In all three plots, the upper and lower branches correspond to stable equilibria, while the middle branch represents an unstable equilibrium, which does not manifest in real-world experiments.

This critical insight of bifurcations remains largely overlooked in the collective intelligence experimental literature. We propose that much of the confusion in the experimental literature can be clarified, and ultimately reconciled, by incorporating the bifurcation phenomenon and its underlying mechanisms into a cohesive theoretical framework. While prior models often rely on specific assumptions that limit their applicability to collective intelligence experiments, we aim to formulate a general, and flexible mathematical model that captures the key effects of social influence on group opinion composition. We first demonstrate that this model reproduces the bifurcation predictions of more complex prior models. We then analyze data from four previous experiments on binary choices, demonstrating that evidence of two equilibria under high social influence is already present in existing data. Using simulations, we replicate the conditions of prior experiments, and show that disparate empirical findings can be predicted and reconciled by the same model under different parameters. Specifically, the prior findings we will replicate are: (1) Initial accuracy non-linearly affects collective accuracy under social influence (Becker et al., 2022); (2) Social influence

Table 1: Summary of theoretical models examining the impact of social influence on collective decision-making in binary choice tasks, reaching bifurcation predictions.

Study	**Agents**	**Task**	**Model method**
Brock and Durlauf (2001)	Human	Make binary choice	Utility maximization
Couzin et al. (2011)	Fish	Choose path to locate food	Agent-based, supported by differential equation extension
Mori et al. (2012)	Human	Make binary choice	Stochastic simulation
Gray et al. (2018)	Bees	Choose nest site	Agent-based
Yang et al. (2021)	Human	Make binary choice	Differential equation, supported by agent-based extensions

can enhance individual accuracy while impairing collective accuracy (Frey & van de Rijt, 2021); and (3) Groups can demonstrate self-correcting dynamics, reducing the prevalence of inferior options in subsequent response (van de Rijt, 2022). Our theory further predicts that some of these conclusions hold only under specific conditions, and we outline the circumstances under which we expect these conclusions to change. We use this model to further delineate conditions under which social influence is expected to improve or hinder collective intelligence.

2 Mathematical model

We consider a group of individuals making a binary choice between options X and Y, which are two potential answers to a factual question, and X is the correct answer. Each individual's decision is modeled as a combination of two components: independent judgment—the choice they would make in isolation—and social influence—the effect of others' choices on their decision. This consideration aligns with several established models of group opinion formation, including the DeGroot (1974) model and the studies reviewed in Table 1.

Independent judgments are shaped by alignment with prior internal beliefs (Dalege et al., 2024). For example, one's opinion on the effectiveness of the COVID-19 vaccine may be closely tied to beliefs about the government and pharmaceutical companies. In this study, we assume that these independent judgments remain stable throughout an experiment. Social influence, on the other hand, often operates on a frequency-dependent basis, particularly in situations where individuals lack clear track records to enable payoff-based learning (Mesoudi, 2016), as is the case in the collective intelligence experiments we are aiming to reconcile. In frequency-dependent social learning, a response that has been demonstrated to be both evolutionarily adaptive and empirically prevalent is the tendency of individuals to disproportionately adopt the majority behavior (Efferson et al., 2008; Henrich & Boyd, 1998). For instance, if 60% of individuals are observed choosing a particular option, the probability of another individual adopting that same option exceeds 60%. In Supplementary Information, we discuss how our predictions would change if this over-response to the majority is not present.

A simple and generalizable way to formulate the combination of individual judgment and social influence mathematically is as follows. For an individual i, we denote the proportion of other individuals choosing X observed by i, as $\tilde{x}_i$. The probability of i choosing the correct option, X, is

$$P(x)_i = (1 - w_i)I + w_i S(\tilde{x}_i) \ . \tag{1}$$

Parameter I, ranging between 0 and 1, denotes the *independent accuracy* for the task—the probability for individuals to choose the correct option independently. It reflects task difficulty, or the degree of alignment between the correct answer and individuals' existing beliefs. A high I value indicates that individuals are likely to answer correctly when acting independently, suggesting the question is relatively easy. When $I = 0.5$, the probability of individuals answering correctly is equivalent to random guessing, indicating a challenging question. For $I < 0.5$, the correct answer is counter-intuitive given existing beliefs, making incorrect responses more likely.

The term $S(\tilde{x}_i)$ denotes the likelihood of choosing an option when the decision is solely based on the observed frequency in others. It follows an S-shaped curve, reflecting the disproportionate adoption of majority behavior characteristic of social learning (Claidière & Whiten, 2012). This function should also satisfy $S(1) = 1$ and $S(0) = 0$, indicating that when an individual makes their decisions solely based on social information, if they observe everyone else choose a particular option, they will also choose that option. Additionally, this function needs to satisfy $S(0.5) = 0.5$, implying that when an individual encounters ambivalent social information, social influence should remain neutral, favoring neither option. One S-shaped functional form satisfying these properties is $S(x) = x^\alpha/(x^\alpha + (1-x)^\alpha)$, which we will use for the remainder of this paper. A visualization of this function is shown in Figure 2(B). We choose this functional form because it is parameterized by a single shape parameter, $\alpha > 1$, allowing us to vary the degree of nonlinearity. In the Supplemental Information, we present analytic results showing the main prediction of the model applies to general S-shaped conformity function and is not sensitive to this choice of functional form.

The parameter w_i represents the extent to which individual i weights social influence relative to their independent judgment, with values ranging from 0 to 1.

2.1 Simulating sequential updating experiments

To assess whether prior experimental results can be reproduced using the mechanisms described above, we integrate these mechanisms with experimental structures. In many experiments and real-world settings, individuals interact sequentially: the first individual makes a choice, the second observes this choice before making their own, the third observes the choices of the first two, and so on. This process can be effectively modeled by simulating Equation 1 within an agent-based simulation. In sequential updating, the strength of social influence starts at zero for the first participant and increases with each subsequent participant in the sequence but with diminishing returns, as social information coming from more individuals is expected to carry more weight. Thus the first individual chooses the correct answer with probability I, and the weight of social information, w, grows as a function of the participant's order in the experiment, i, with diminishing returns. This can be mathematically formulated as $w_i = a\ i/(1 + a\ i)$, where a is a positive constant. The first participant has order $i = 0$. A visualization of this function is shown in Figure 2(C). The choice of this functional form is motivated by its use in the Polya Urn model (Mahmoud, 2008), a classic framework for path-dependent phenomena. Greater values of a imply stronger social learning. Note that when $i = 1/a$, $w_i = 1/2$. This implies that $1/a$ represents the number of individuals whose influence is weighted equally to an individual's own independent judgment. In the sequential updating formulation, $\tilde{x}_i$ in Equation 1 denotes the proportion of participants choosing option X prior to participant i.

2.2 Simulating synchronous updating experiments

The second type of process studied in experiments and observed in real-world scenarios involves individuals initially forming opinions independently, announcing them synchronously, followed by several rounds of interaction where individuals update their positions, culminating in a majority vote. In this type of process, heterogeneity in social influence strength can be ignored, setting $w_i = w$ for all individuals. Assuming all-to-all interaction, as is the case in the experiments we aim to reproduce, the observed proportion choosing option X is the overall proportion in the group, x. Thus, Equation 1 simplifies to,

$$P(x)_i = (1 - w)I + wS(x) \ . \tag{2}$$

We will use an agent-based simulation of Equation 2, where agents update simultaneously to simulate one round of synchronous update. Updates are performed multiple times, to generate the predicted distribution of responses in the synchronous updating experiment.

Note that the individual-level dynamics described by Equation 2 lends itself to a simple mean-field formulation at the group level in the limit of large group sizes. At any time t, the proportion of individuals choosing X is given by $x_{\text{ind}} = (1 - w)I + wS(x)$. Assuming individuals update their choices on a timescale τ, the evolution of the group's proportion selecting option X can be described by the following differential equation,

$$\frac{dx}{dt} = \frac{x_{ind} - x}{\tau} \tag{3}$$

We will also analyze the equilibrium solutions of this differential equation, which satisfy $x = (1 - w)I + wS(x)$, to derive theoretical results.

3 Results

3.1 Model's bifurcation predictions

This model, in both the sequential updating and synchronous updating forms, reproduces the bifurcation observed in prior theoretical literature using more complex models. Figure 3(A) presents results using the sequential updating version of the model, derived via agent-based simulations of Equation 1. Each marker represents a simulation of a group of 100 individuals, with 50 groups simulated for each level of social influence. The results demonstrate that under high social influence, some groups overwhelmingly select the correct answer, while others converge on the incorrect one.

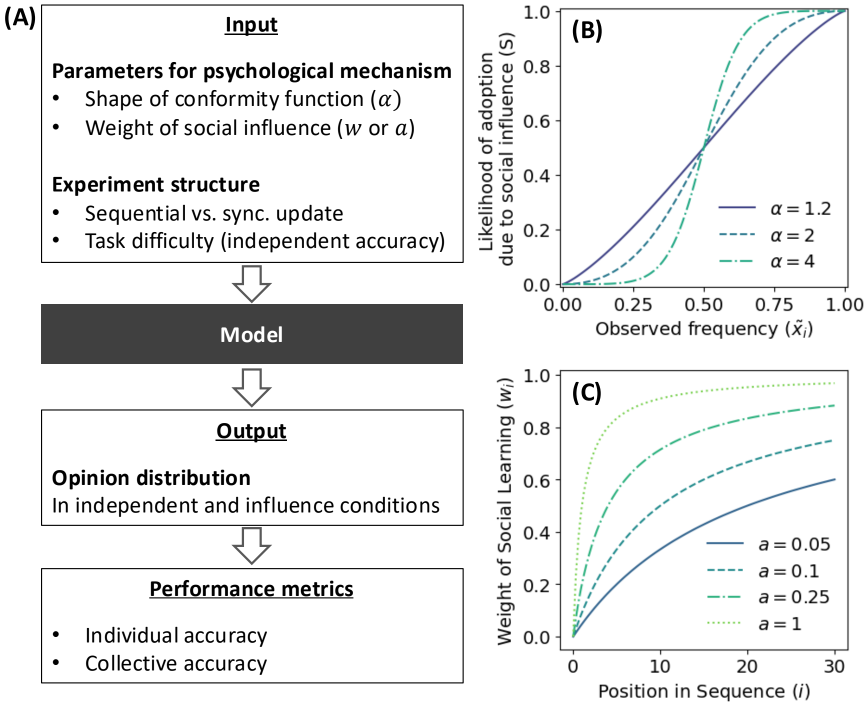
Similarly, Figure 3(B) shows predictions for synchronous updating experiments. The dots represent results from agent-based simulation, where each marker represents a group of 100 individuals, and each group had undergone 100 iterations of update, and x has reached convergence. Also 50 groups are simulated for each level of social influence. We also show the equilibrium solution based on the differential equation (Equation 3). The solid lines represent the solutions to stable equilibria, while the dashed line indicates unstable equilibria. Unstable equilibria are not expected to be observed in real-world experiments, as any fluctuation or noise would cause the group to deviate from them and move to a stable equilibrium. Instead, these unstable equilibria serve as boundaries delineating the basins of attraction for the two stable equilibria. In both the agent-based simulation and equilibrium solutions, a similar bifurcation appears in the synchronous case. However, the prevalence of the low-accuracy equilibrium is less prevalent than in the sequential updating scenario. In simulations of both panels, and throughout the manuscript, we use shape parameter for the conformity function $\alpha = 2$.

The bifurcation results are those found in prior models such as in Figure 2. The intuition behind these results is as follows: at one extreme, when individuals act independently, the group's split reflects the task's independent accuracy. At the other extreme, when individuals rely solely on social information, the group converges entirely on one option, which could be either correct or incorrect. Thus, a bifurcation must occur in the transition between these extremes, shifting from one outcome to two.

3.2 Bifurcation in existing experimental data and replicating experiments in models

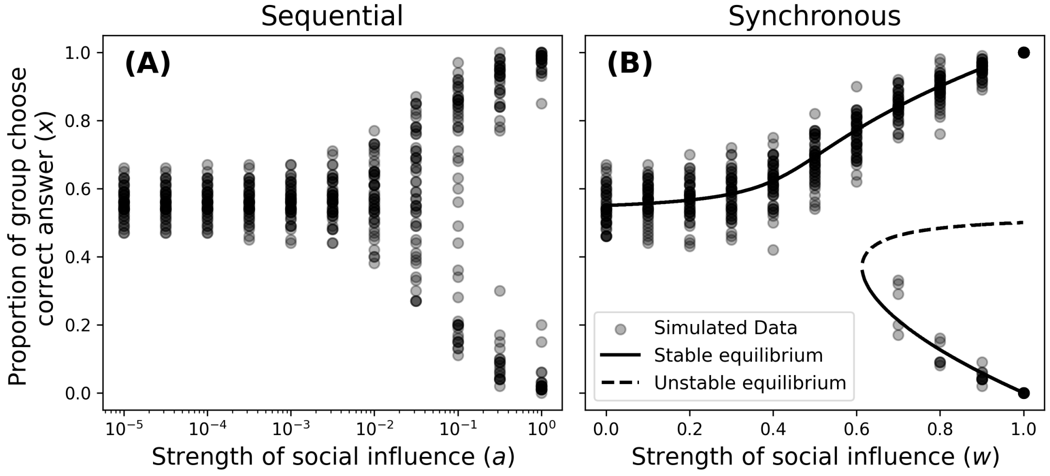
The model's bifurcation prediction suggests that, under high social influence, the proportion of the group choosing either option should exhibit a bimodal distribution, reflecting noisy realizations

Figure 2: (A) Summary of the model input and output used to reconcile desperate experimental findings. (B, C) Plots of functional forms used in the model.



Notes. (B) $S(x)$, representing frequency-based conformity response, where likelihood of adoption increases with observed frequency. (C) w_i for sequential updating experiments, illustrating the weights of social information in sequential updating experiments, which increases with the individual's position in the sequence.

Figure 3: Model prediction for the proportion of individuals in each group that choose the correct outcome, replicating bifurcation prediction in prior literature.



Notes. (A) For sequential updating experiments, based on agent-based simulation. Each marker represents a simulation of one group. (B) For synchronous updating experiments, dots are based on agent-based simulation. Solid lines and dashed line represent stable and unstable equilibria of the mean-field differential equation for the system (Equation 3).

of the two equilibria. In contrast, under independent conditions, the model predicts an unimodal distribution, as only one equilibrium is expected. To test these predictions, we reanalyze data from four prior experiments, Frey and van de Rijt (2021)'s laboratory and online experiments, Becker et al. (2022)'s binary exchange experiment, and the experiment in Mori et al. (2012) (see Supplementary Information for details on the data). All of the experiments are for binary choice tasks over factual questions. They all compare the outcomes of an independent condition, where individuals report their choices independently, with outcomes of an influence condition, where individuals interact with one another. Three experiments (Frey & van de Rijt, 2021; Mori et al., 2012) use sequential updating, and one (Becker et al., 2022) uses synchronous updating. Since the model predictions are for tasks with a single level of independent accuracy, and the experiments include questions with varying independent accuracy, we ensure a valid comparison by slicing the experimental data for questions with independent accuracy between 0.4 and 0.6. This range, corresponding to participants having a 40% to 60% chance of independently selecting the correct option, also contains the densest data in the datasets.

Figure 4's left column shows the distributions of the proportion of individuals that chose the correct option in each group in the four experiments analyzed, with each observation representing a group. The influence conditions exhibit a wider spread than the independent conditions. This suggests that social influence leads to greater uncertainty in the group's outcome, which is qualitatively consistent with our model's predictions.

We replicate the distributions in the social influence conditions of these experiments using agent-based simulations of our model, taking as input the same distribution of independent accuracy and group size as in the original studies. Across all simulations, we set the parameter $\alpha = 2$. For in-person sequential updating experiments (Frey and van de Rijt (2021) lab and Mori et al. (2012)), we use $a = 0.33$; for the online sequential updating experiment (Frey and van de Rijt (2021) online), we use $a = 0.08$; and for the synchronous experiment (Becker et al. (2022)), we use $w = 0.75$. The synchronous experiment involved only two rounds of updates. The middle column of Figure 4 shows our simulation results, which qualitatively align with the empirical data.

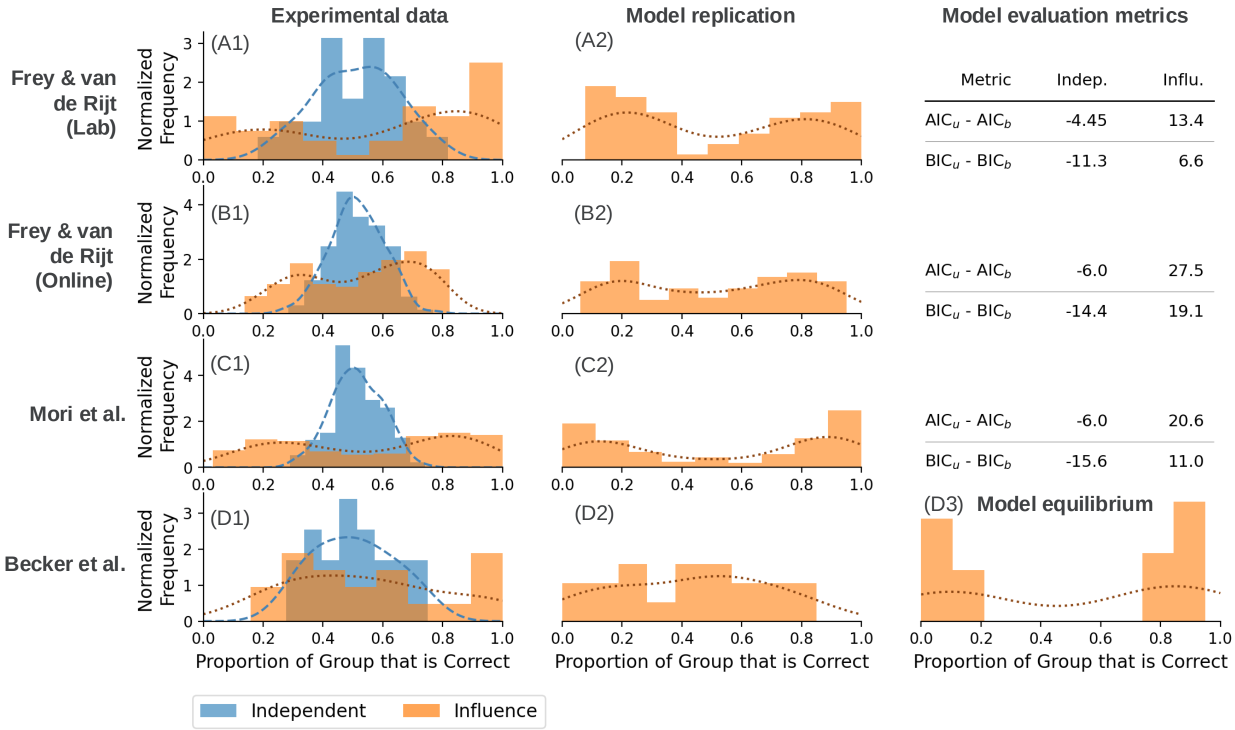
We quantitatively evaluate whether the experimental data are unimodal or bimodal as follows. We fit these distributions to a truncated Gaussian, truncated between 0 and 1 (since proportions cannot exceed these bounds), representing the unimodal fit, and to the normalized sum of two truncated Gaussians, representing the bimodal fit. We use the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to evaluate which model better fits the data. These metrics assess how well a model fits the data while penalizing models with more parameters, with BIC applying a heavier penalty. Lower AIC and BIC values indicate a better-fitting model. Thus if an unimodal fit better describes the data, than a bimodal fit, we would expect $\mathrm{AIC}_u - \mathrm{AIC}_b < 0$ and $\mathrm{BIC}_u - \mathrm{BIC}_b < 0$, where AIC_u and BIC_u are the AIC and BIC for the unimodal fit, and AIC_b and BIC_b are those for the bimodal fit. Conversely, if the bimodal fit is better than the unimodal fit, we would expect $\mathrm{AIC}_1 - \mathrm{AIC}_2 > 0$ and $\mathrm{BIC}_1 - \mathrm{BIC}_2 > 0$. The right side of Figure 4 presents the model evaluation metrics for the three sequential updating experiments. In all cases, these metrics support a bimodal distribution for the influence condition and an unimodal distribution for the independent condition. The one exception is the Becker et al. (2022) experiment, where the metrics indicate an unimodal fit for both conditions (Independent condition: $\mathrm{AIC}_u - \mathrm{AIC}_b = -3.98$, $\mathrm{BIC}_u - \mathrm{BIC}_b = -6.97$; Influence condition: $\mathrm{AIC}_u - \mathrm{AIC}_b = -1.55$, $\mathrm{BIC}_u - \mathrm{BIC}_b = -4.54$). We attribute this to the experiment's synchronous updating being limited to only two iterations. Our model replicates this pattern when restricted to two iterations (Figure 4(D2)), as strong bimodality does not emerge within such a short time frame. However, when we extend the simulation to 100 iterations—where x converges—the model predicts a bimodal distribution (Figure 4(D3)).

In sum, bimodal distributions—resulting from the existence of two equilibria—are observed in the social influence conditions of some, but not all, studies. This suggests that group majority may arrive at an outcome in a given experiment but could reach a different outcome if the same experiment is repeated.

3.3 Initial accuracy positively associated with collective accuracy after social influence

A finding reported in Becker et al. (2022) is that initial estimates significantly impact the effect of social influence on collective accuracy. For groups that have a high accuracy before social influence, social influence further improves group accuracy. For those groups that have a low accuracy before social influence, social influence decreases group accuracy. The effect diminishes for initial accuracy

Figure 4: Distribution of the proportion of individuals in each group choosing the correct option across four existing experiments compared with model's predictions.



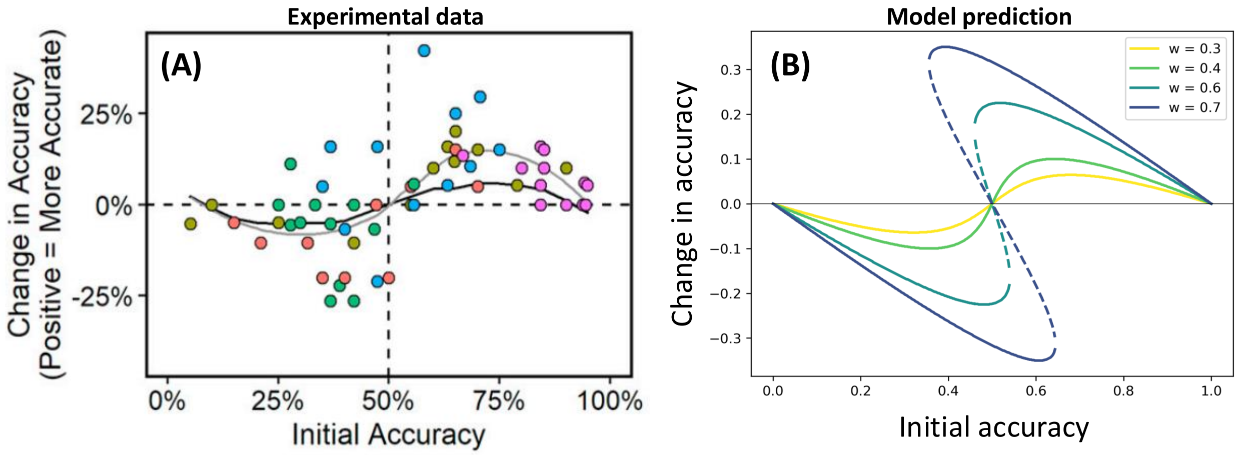
Metric	Indep.	Influ.
AIC_u - AIC_b	-4.45	13.4
BIC_u - BIC_b	-11.3	6.6
AIC_u - AIC_b	-6.0	27.5
BIC_u - BIC_b	-14.4	19.1
AIC_u - AIC_b	-6.0	20.6
BIC_u - BIC_b	-15.6	11.0

Notes. Independent conditions are shown in blue, while social influence conditions are shown in orange. Dashed lines represent kernel density estimates of these distributions. The left column presents experimental data, the middle column displays the model's replication of the experiment, and the right column (for the first three experiments) reports model evaluation metrics, all of which support bimodal fits for influence condition and unimodal fit for independent condition. In the last row, the data and replication support unimodal fits for both conditions. Panel (D3) shows the model replication ran to equilibrium, illustrating that the absence of bimodal behavior may stem from the synchronous updating experiment analyzed being limited to only two rounds of updating.

of 0%, 50%, and 100%. These findings are shown in Figure 5(A), where positive values indicate that the social influence condition has higher accuracy.

Since this experiment uses synchronous updating, we replicate these results by simulating the synchronous updating version of our model across varying levels of independent accuracy. We then compare the change in collective accuracy—defined as the proportion of groups whose majority selects the correct option—between the influence and independent conditions. The results, shown in Figure 5(B), illustrate the relationship between initial accuracy (which is the same as independent accuracy for synchronous updating) and change in collective accuracy between the two conditions for various social influence weights (w). The solid lines show stable equilibria of the model, and dashed lines show unstable equilibria, which would not be observed in an experiment. The pattern seen in the prior experiment agrees with the simulated outcome for smaller values of w.

Figure 5: Data from Becker et al. (2022) on the relationship between initial accuracy and whether social influence improves collective accuracy (A) compared with our model's prediction (B).



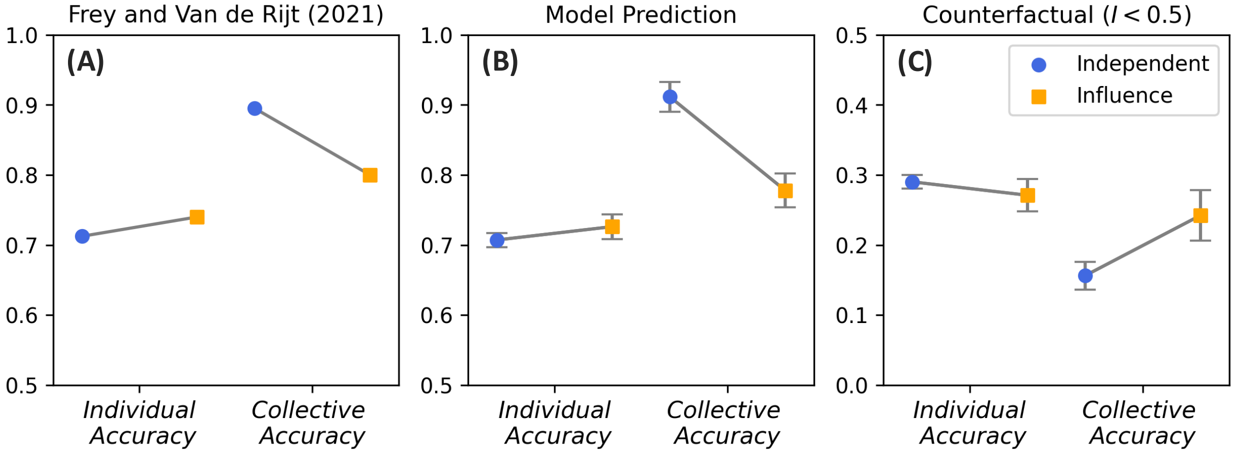
Notes. (B) Shows our model's prediction for different levels of social influence (w), recovering the same empirical patterns as (A), that social influence improves accuracy for initially accurate groups, but hurts accuracy for initially inaccurate groups. Solid lines represent stable equilibria, while dashed lines represent unstable equilibria, which would not be observed in experiments.

3.4 Social influence improves individual accuracy and reduces collective accuracy

A finding reported in Frey and van de Rijt (2021) is that social influence improves individual accuracy, the likelihood of any individual choosing the correct answer, while reducing collective accuracy, the likelihood for the majority of groups to choose correctly. Data from this experimental study are visualized in Figure 6(A). Since this experiment is performed with sequential updating, we reproduce this finding by simulating the sequential updating version of the mathematical model (Equation 1) in agent-based simulations. The simulations take as input the same distribution of independent accuracy (I) and group sizes as the original experiments. Each condition is simulated 30 times, and the results, shown in Figure 6(B), confirm the reported findings: individual accuracy improves under social influence, while collective accuracy decreases ($p < 0.001$).

However, this effect is only expected to hold for tasks with independent accuracy greater than chance ($I > 0.5$). The analysis in Frey and van de Rijt (2021) excludes tasks with $I < 0.5$. As noted in their supplementary information, the effect is predicted to reverse for tasks with $I < 0.5$. We repeat the simulations using independent accuracy values reversed from the original experiment ($1 - I$), resulting in a distribution of independent accuracy values all below 0.5. As shown in Figure 6(C), the reversal is reproduced by our model: social influence reduces individual accuracy while improving collective accuracy ($p < 0.001$).

Figure 6: Comparing data of Frey and van de Rijt (2021) with model's predictions.



Notes. (A) Mean values from experimental data of Frey and van de Rijt (2021), showing that social influence improves individual accuracy but reduces collective accuracy. (B) Simulation results replicating the experiment using the mathematical model and the same task difficulty, reproducing effects observed in the experiments. (C) Simulation results for a counterfactual scenario with lower independent accuracy ($I < 0.5$), showing reversed effects where social influence harms individual accuracy but benefits collective accuracy. In (B) and (C), the model was simulated 30 times for each condition, with error bars representing the standard deviation across simulations. For all simulation results, $p < 0.001$.

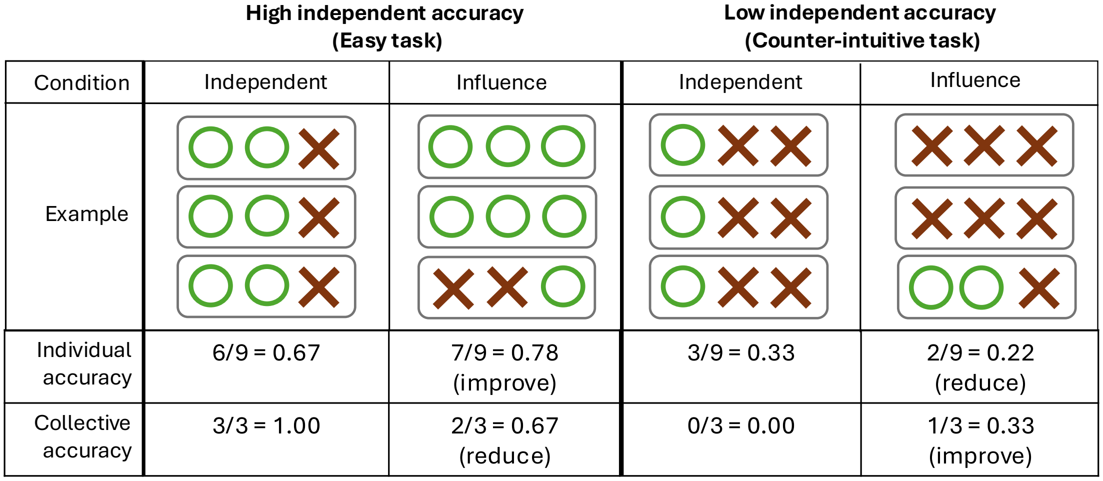
Our theoretical model provides a framework for understanding these results intuitively. We illustrate using a simplified example of three groups, each consisting of three individuals, as shown in Figure 7. For tasks with high independent accuracy (i.e., easier tasks), social influence tends to improve individual accuracy by amplifying the majority's correct position. However, this improvement primarily occurs in groups where the majority would have been correct even without influence. Combined with the occasional amplification of an incorrect minority view, in the less likely event that the incorrect minority speaks first, this dynamic can reduce collective accuracy. This is illustrated in the first two columns of Figure 7: most of the time, social influence boosts the frequency of the correct option (green circle), but this tends to happen in groups that would have been correct without influence (first two rows). Occasionally, if an incorrect minority speaks first, subsequent individuals may adopt this position, flipping a correct majority into an incorrect one and reducing collective accuracy (last row). In this example, individual accuracy is six out of nine (0.67) in the independent condition and increases to seven out of nine (0.78) in the influence condition, indicating an improvement in individual accuracy. However, collective accuracy is three out of three (1.00) in the independent condition, as the majority in all groups select the correct answer. In the influence condition, the correct individuals are concentrated in groups that would have been correct independently. In the last group, the group's majority choose incorrectly. As a result, collective accuracy decreases to two out of three (0.67) in the influence condition. Social influence leads to what resembles a "gerrymandering" of choices, where the correct answer is concentrated within groups that would have been correct without social influence.

Similar dynamics occur for counter-intuitive tasks with independent accuracy of less than 0.5, illustrated in the last two columns of Figure 7. Here, social influence tends to amplify the majority's incorrect opinion, reducing individual accuracy. However, because this decline usually occurs in groups that are already incorrect independently, and due to the occasional ability of a correct minority to speak first and sway the group, social influence can increase collective accuracy.

3.5 Group self-correcting dynamics

Prior research has shown that groups often exhibit self-correcting dynamics—when an inferior option gains popularity, the group adjusts by choosing the inferior option at a frequency lower than its current relative popularity. This dynamic prevents lock-in, a scenario where choices dispropor-

Figure 7: An illustration of how social influence can improve individual accuracy while reducing collective accuracy for tasks with high independent accuracy and have the opposite effect for tasks with low independent accuracy.



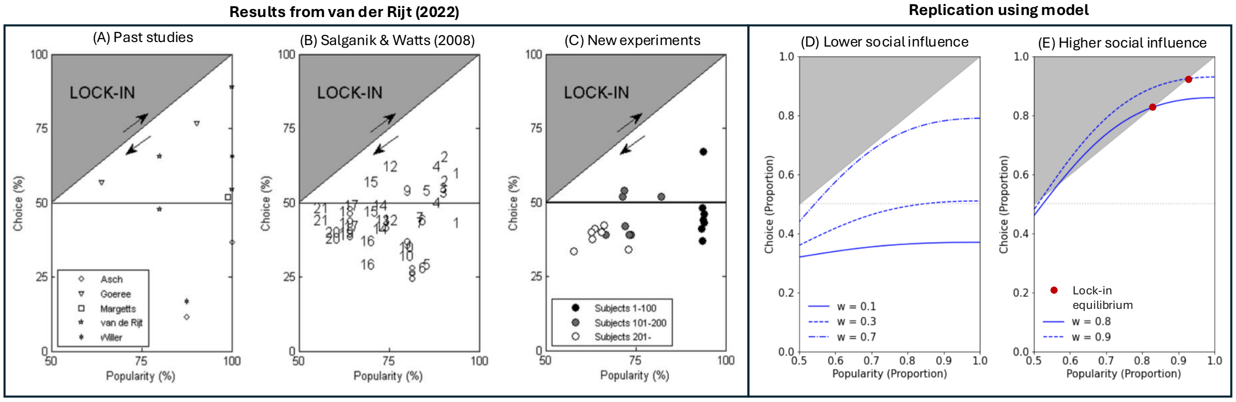
	High independent accuracy (Easy task)		Low independent accuracy (Counter-intuitive task)	
Condition	Independent	Influence	Independent	Influence
Example	○○✗ ○○✗ ○○✗	○○○ ○○○ ✗✗○	○✗✗ ○✗✗ ○✗✗	✗✗✗ ✗✗✗ ○○✗
Individual accuracy	6/9 = 0.67	7/9 = 0.78 (improve)	3/9 = 0.33	2/9 = 0.22 (reduce)
Collective accuracy	3/3 = 1.00	2/3 = 0.67 (reduce)	0/3 = 0.00	1/3 = 0.33 (improve)

Notes. Demonstrated using the example of three groups, each consisting of three individuals. Green circles denote correct answers and red crosses denote incorrect answers.

tionately align with popularity, allowing an inferior option to dominate, which corresponds to the grey-shaded regions of Figure 8. Above the diagonal line, the frequency of selection exceeds the current popularity, reinforcing dominance and locking the group into the inferior choice. Reanalysis of several behavioral experiments and an original experiment van de Rijt (2022) demonstrate groups' self-correcting dynamics. These span contexts including the Asch conformity experiment, crowdfunding for projects, product reviews, and wine-wasting show lock-ins are avoided, shown in Figure 8(A). Figure 8(B) presents reanalysis of the music lab experiment Salganik and Watts (2008), which demonstrates the unpredictability of which songs achieve success due to information cascades of initial popularity. In this figure, each number corresponds to a pair of songs. Figure 8(C) displays results from the original experiment, where participants are tasked with making binary choices. Across all datasets, self-correcting behavior is evident, as every data point remains below the lock-in threshold, represented by the diagonal line.

It is important to note that self-correction is not guaranteed. Many social processes exhibit lock-in effects, such as the dominance of technologies with network effects (Arthur, 1989), the persistence of bestsellers and celebrity status (van de Rijt, 2022), and the outcomes of some collective intelligence experiments (Frey & van de Rijt, 2021).

Figure 8: Results from van de Rijt (2022) compared with model predictions



Notes. (A)–(C) Figures from van de Rijt (2022) showing the frequency of choosing the inferior alternative as a function of its popularity across various experiments. In all cases, the data remain outside the lock-in region, where the popularity of the inferior alternative would be further amplified by individual's choices. (D) Model prediction showing probability of choosing an option based on its frequency for different levels of social influence (w). The model replicates the experimental findings that lock-in is avoided when social influence is below a threshold. (E) At higher levels of social influence, lock-in becomes possible. The red dots mark the points at which lock-in is predicted to occur, where the popularity equals the choice probability.

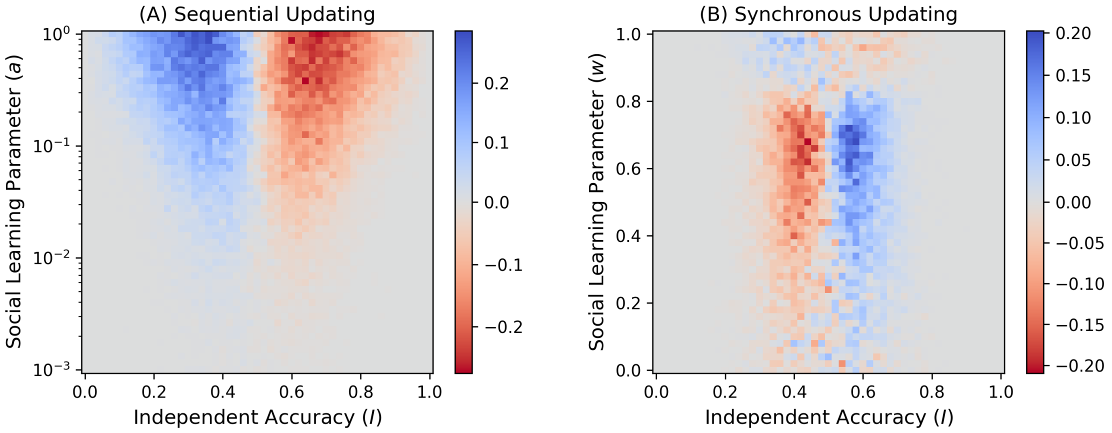
Self-correcting behavior and the possibility of not self-correcting can both be explained with our model by variation of the weight of the social influence parameter, w. With the simplifying assumption that all individuals have the same w value, the proportion of individuals choosing an option after social influence, given its current popularity x, is $P(X) = (1 - w)I + wS(x)$. While the model was formulated assuming X is the correct option, and x is the proportion choosing the correct option, a similar equation can be formulated for the inferior option, Y, whose popularity (in proportions) is y, $P(Y) = (1-w)I_y + wS(y)$, where I_y is the likelihood of choosing Y independently. We plot this relationship between choice probability ($P(Y)$) and popularity (y) for various values of w, with results shown in Figure 8(D) and (E). In these simulations, we set $I_y = 0.3$, meaning that when individuals make independent choices, 30% will select the inferior option. When the level of social influence is below a certain threshold, the popularity-choice curve remains below the lock-in region (panel D). As social influence increases further, the curve can enter the lock-in region, as shown in panel E. The red dots indicate the points where popularity equals choice probability (where the curve crosses the diagonal line), representing stable equilibrium positions of the group composition. Once lock-in occurs, the inferior option dominates the majority, though a minority group resists adopting the inferior option. Thus we show self-correcting dynamics occur for some values in the parameter space, but it is not guaranteed. These experimental results are a dynamic interplay between individual judgments and social influence, with the extent of social influence modulating whether self-correcting dynamics occur.

3.6 Delineating conditions under which social influence harms or improves collective accuracy

In the sections above, we show several conclusions from prior experiments can be explained by integrating the same mathematical model with their respective experimental procedures. Here, we use the model to predict general conditions under which social influence is expected to improve or hinder collective accuracy. Using our model, we compute the predicted difference in collective accuracy—defined as the probability of the group majority choosing correctly (Prob($x > 0.5$))—between the social influence and independent conditions, where positive values indicate social influence condition being more accurate. These results, shown in Figure 9, are presented as a function of independent accuracy, which reflects task difficulty, and the degree of social influence, which is controlled by parameter a for the sequential updating version of our model and by parameter w for the synchronous updating version. For each point in the parameter space (a 51×51 grid), we simulate 400 groups of 21 individuals and compare the average collective accuracy between social influence and independent conditions. The synchronous condition is simulated to equilibrium in x with 75 revisions. Blue regions indicate where social influence enhances collective accuracy, red regions indicate where it diminishes accuracy, and grey regions indicate no predicted effect. In the Supplementary Information, we show that varying group size does not qualitatively alter these results, though it may affect the size of the blue and red areas.

The results reveal a fundamental distinction between sequential (Figure 9(A)) and synchronous (Figure 9(B)) updating processes. For easier tasks (independent accuracy $I > 0.5$), social influence reduces collective accuracy in sequential updating. In synchronous updating, a similar effect occurs at very high levels of social learning; however, at more moderate levels, the relationship reverses, and social learning enhances collective accuracy. For counterintuitive tasks ($I < 0.5$), these patterns are reversed: social influence improves collective accuracy in sequential updating, whereas in synchronous updating, this improvement is observed only at very high levels of social learning, while moderate levels lead to a decline in accuracy.

Figure 9: Model's predicted difference in collective accuracy between social influence and independent conditions.


(A) Sequential Updating

(B) Synchronous Updating

Notes. (A) For sequential updating experiments. (B) For synchronous updating experiments. Blue regions (positive values) indicate where social influence is expected to improve collective accuracy compared to independent conditions, while red regions (negative values) indicate where it is expected to reduce collective accuracy. For both panels, the vertical axis is a model parameter where high values indicate greater weight of social learning. For the same level of independent accuracy, social learning can have the opposite effect on collective accuracy depending on whether the experiment is performed using sequential or synchronous updating.

While both updating processes produce similar bifurcation patterns (Figure 3), they lead to different collective accuracy outcomes. This difference arises because the probability of reaching high- or low-accuracy equilibria varies between the two processes. Compared to synchronous updating, sequential updating is more likely to result in low-accuracy equilibria, as outcomes heavily depend on the accuracy of the initial decision-maker. An incorrect first decision significantly increases the likelihood of the group converging on the low-accuracy outcome. In contrast, synchronous updating requires a larger proportion of individuals to be initially incorrect for the group to shift toward the low-accuracy equilibrium, making such outcomes less probable. This distinction is evident in Figure 3, where the prevalence of simulations clustering around low-accuracy equilibria is greater in (A) than in (B).

The model's prediction that sequential and synchronous updating leads to different conclusions aligns with prior findings. For instance, easier tasks are associated with higher initial accuracy. According to Becker et al. (2022), social influence tends to enhance collective accuracy in such tasks. However, Frey and van de Rijt (2021) also examine easy tasks and find that social influence reduces collective accuracy. This apparent contradiction can be reconciled by taking into account the fact that the former study employs synchronous updating, while the latter uses sequential updating, and we should expect social influence to have the opposite effect on collective accuracy in these two processes for a moderate level of social learning. The model's prediction also suggests that task difficulty plays a crucial role in determining whether social influence helps or harms collective accuracy, as indicated by the reversal of results in Figure 9 across the $I = 0.5$ line. This finding aligns with the results in Section 3.4, where social influence has opposite effects on collective accuracy for easy versus counterintuitive tasks.

4 Discussion

We have demonstrated that seemingly disparate—and at times contradictory—experimental findings can be effectively reconciled within a unified mathematical framework that models how individuals integrate independent judgments and social information, when combined with specific experimental designs. The model predicts that in easy tasks, a very high level of social influence is always detrimental. However, for a moderate level of social influence, the impact depends on the mode of updating: synchronous updating enhances collective accuracy, while sequential updating undermines it. Task difficulty also plays a crucial role—these relationships reverse for counter-intuitive tasks, where the correct answer contradicts prior beliefs.

These findings suggest that for everyday decisions where independent judgments are typically reliable, it is crucial to create mechanisms that allow individuals to form and express their opinions independently rather than being influenced by others before forming their own views. Practical strategies to mitigate the downsides of sequential updating include using real-time online polling tools to gather and share responses simultaneously or encouraging group members to independently think through and write down their answers before sharing them with the group, such as the Delphi method. These approaches help preserve the benefits of social influence while avoiding the pitfalls of premature influence.

Researchers study collective intelligence typically with the goal of determining how to structure groups for optimal performance. Our findings suggest that when a group faces uncertainty——such as not knowing in advance whether a task is easy or counterintuitive——designing the optimal communication structure becomes challenging. The same structure can yield opposite effects depending on task difficulty. This underscores the need to move beyond a purely optimization-focused approach in collective intelligence research. Instead of seeking a single "best" structure, it is crucial to consider how groups can adapt to a broader range of environments (Galesic et al., 2023). Effective collective intelligence may require groups to dynamically adjust their social structures as they encounter new tasks.

Our model makes several simplifying assumptions to maintain parsimony. We assume that the weight of social influence does not change with the level of consensus among others. While our model accounts for individuals being more inclined to choose an option when it is overwhelmingly favored (as captured in the formulation of the S function), we acknowledge that, in reality, the weight of social influence (w) may also increase as consensus strengthens. Additionally, our model neglects the effect of having a single ally in social influence—individuals are known to conform significantly less when they have one ally compared to none (Morris & Miller, 1975). Since our model formulates social influence in terms of proportions, it does not capture this effect. Future work could enhance our model by incorporating these more complex hypotheses to better reflect real-world dynamics.

In this paper, we focus on binary choice tasks with a known correct answer. However, many real-world decisions do not have known correct answers at the time the decisions are made. Examples include determining which strategy a company should adopt, whether legislators should pass a particular bill. While a factually correct answer does not exist when the group makes a choice, certain choices may yield more adaptive outcomes—though only apparent in hindsight. For example, one strategy can lead to greater future revenue for the company, and one bill aimed at reducing the unemployment rate turns out more successful than the alternative. From the perspective of individual participants, however, the process is the same whether or not a correct answer exists in priori or it is only known in hindsight, as they remain agnostic to the correct answer throughout the collective decision-making process.

Our model can be further extended to scenarios where no objective right or wrong answer exists, even in hindsight. Such situations are particularly relevant to democratic elections, such as deciding which candidate should be elected president. In these cases, a sensible objective for the collective decision making process is for collective decisions to reflect the majority preference among participants' independent judgments. In these cases, the parameter I can be reinterpreted as the likelihood of any individual to prefer X over Y while independent. Within this new interpretation of parameters, the model could be used to evaluate whether the collective decision aligns with the majority's independent preferences. Assuming $I > 0.5$, that is effectively selecting option X.

While our model focuses on binary choices, most experiments in collective intelligence involve numerical estimation tasks, likely inspired by Galton (1907)'s classic demonstration of cow weight estimation. However, many of the highest-stakes decisions in organizations and society are discrete choices—such as selecting a president or strategy. This consideration motivated our choice to focus on discrete choices in our study, and we advocate for more experimental research on such decisions within the collective intelligence community. Although discrete choice and numerical estimation tasks are closely related—for instance, some discrete decisions can be derived by thresholding continuous values—Becker et al. (2022) demonstrated that insights from numerical estimation tasks cannot always be directly applied to related discrete choices derived from thresholding. Future studies could also extend our modeling framework to numerical estimation tasks by replacing the social influence term, $S(\tilde{x})$, with a function that incorporates the mean or median of others' numerical estimates.

In collective intelligence research and social behavioral sciences more broadly, there have been serious concerns over experiments conducted under narrowly specified conditions attempting to make broad conclusions that contradict each other. The condition of applicability for these experimental results is frequently under-specified, complicating efforts to derive actionable insights. Our finding underscores this issue, showing that aspects of experimental design beyond the primary independent variable—social influence—such as sequential versus synchronous updating and task difficulty, play a critical role in shaping conclusions about whether social influence improves or impairs collective accuracy.

The same concern over the applicability problem has been brought to attention in the collective intelligence community by Almaatouq et al. (2024). Echoing Newell (1973), they liken such experiments to "playing 20 questions with nature," oversimplifying complex phenomena into a series of binary answers. To address this concern, they advocate for using machine learning to predict general outcomes from experimental data, moving away from human-centered interpretations in order to gain predictability. While we acknowledge the same problem, we believe the solution need not forgo interpretability. Our results demonstrate that by carefully formulating simple, human-interpretable mathematical models, we can reconcile disparate findings. Despite the complexity of the experimental literature, many outcomes can be understood as the interaction of the same psychosocial mechanisms of human behavior with various structure of experimental setups. The insights derived from these models can help the design of future collective intelligence experiments to carefully consider nuanced but critical factors, such as which updating procedure and tasks to use.

Besides implications for the collective intelligence community, our work raises important questions for the broader social and behavioral research community. The multi-equilibrium behavior our work highlights is a common feature of collective human behavior driven by reinforcing feedback loops that amplify initial advantages (Arthur, 1989; Sterman, 2000). How can experiments meaningfully account for the complexity of multi-equilibrium dynamics? Most experimental methods rely on comparing means between conditions, but this approach fails to capture the important properties of bimodal distributions effectively. This underscores the need for better methods to study multi-equilibrium phenomena. Inspiration may come from "multiple-worlds" experiments, such as the music lab study by Salganik et al. (2006) or the partisan polarization experiment by Macy et al. (2019), which explicitly examine the existence and characteristics of multiple equilibria.

Acknowledgments

This research was partially supported by National Science Foundation award 2329988.

References

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2024). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, *47*, e33.

Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *117*(21), 11379–11386.

Almaatouq, A., Rahimian, M. A., Burton, J. W., & Alhajri, A. (2022). The distribution of initial estimates moderates the effect of social influence on the wisdom of the crowd. *Scientific Reports*, *12*(1), 16546.

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal*, *99*(394), 116–131.

Becker, J. A., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *114*(26), E5070–E5076.

Becker, J. A., Guilbeault, D., & Smith, E. B. (2022). The crowd classification problem: Social dynamics of binary-choice accuracy. *Management Science*, *68*(5), 3949–3965.

Brock, W. A., & Durlauf, S. N. (2001). Discrete choice with social interactions. *The Review of Economic Studies*, *68*(2), 235–260.

Claidière, N., & Whiten, A. (2012). Integrating the study of conformity and culture in humans and nonhuman animals. *Psychological Bulletin*, *138*(1), 126.

Couzin, I. D., Ioannou, C. C., Demirel, G., Gross, T., Torney, C. J., Hartnett, A., Conradt, L., Levin, S. A., & Leonard, N. E. (2011). Uninformed individuals promote democratic consensus in animal groups. *Science*, *334*(6062), 1578–1580.

Da, Z., & Huang, X. (2020). Harnessing the wisdom of crowds. *Management Science*, *66*(5), 1847–1867.

Dalege, J., Galesic, M., & Olsson, H. (2024). Networks of beliefs: An integrative theory of individual- and social-level belief dynamics. *Psychological Review*.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*(345), 118–121.

Efferson, C., Lalive, R., Richerson, P. J., McElreath, R., & Lubell, M. (2008). Conformists and mavericks: The empirics of frequency-dependent cultural transmission. *Evolution and Human Behavior*, *29*(1), 56–64.

Farrell, S. (2011). Social influence benefits the wisdom of individuals in the crowd. *Proceedings of the National Academy of Sciences*, *108*(36), E625–E625.

Frey, V., & van de Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science*, *67*(7), 4273–4286.

Galesic, M., Barkoczi, D., Berdahl, A. M., Biro, D., Carbone, G., Giannoccaro, I., Goldstone, R. L., Gonzalez, C., Kandler, A., Kao, A. B., et al. (2023). Beyond collective intelligence: Collective adaptation. *Journal of the Royal Society interface*, *20*(200), 20220736.

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Gray, R., Franci, A., Srivastava, V., & Leonard, N. E. (2018). Multiagent decision-making dynamics inspired by honeybees. *IEEE Transactions on Control of Network Systems*, *5*(2), 793–806.

Gürçay, B., Mellers, B. A., & Baron, J. (2015). The power of social influence on estimation accuracy. *Journal of Behavioral Decision Making*, *28*(3), 250–261.

Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior*, *19*(4), 215–241.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389.

Jayles, B., Kim, H.-r., Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., Sire, C., & Theraulaz, G. (2017). How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, *114*(47), 12620–12625.

Kao, A. B., Berdahl, A. M., Hartnett, A. T., Lutz, M. J., Bak-Coleman, J. B., Ioannou, C. C., Giam, X., & Couzin, I. D. (2018). Counteracting estimation bias and social influence to improve the wisdom of crowds. *Journal of The Royal Society Interface*, *15*(141), 20180130.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025.

Macy, M., Deri, S., Ruch, A., & Tong, N. (2019). Opinion cascades and the unpredictability of partisan polarization. *Science Advances*, *5*(8), eaax0754.

Mahmoud, H. (2008). *Pólya urn models*. Chapman; Hall/CRC.

Mesoudi, A. (2016). Cultural evolution: A review of theory, findings and controversies. *Evolutionary Biology*, *43*, 481–497.

Minson, J. A., Mueller, J. S., & Larrick, R. P. (2018). The contingent wisdom of dyads: When discussion enhances vs. undermines the accuracy of collaborative judgments. *Management Science*, *64*(9), 4177–4192.

Mori, S., Hisakado, M., & Takahashi, T. (2012). Phase transition to a two-peak phase in an information-cascade voting experiment. *Physical Review E*, *86*(2), 026109.

Morris, W. N., & Miller, R. S. (1975). The effects of consensus-breaking and consensus-preempting partners on reduction of conformity. *Journal of Experimental Social Psychology*, *11*(3), 215–223.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press,

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854–856.

Salganik, M. J., & Watts, D. J. (2008). Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly*, *71*(4), 338–355.

Sterman, J. (2000). *Business dynamics*. Irwin/McGraw-Hill c2000..

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

van de Rijt, A. (2022). Self-correcting dynamics in social influence processes. In *Handbook of sociological science* (pp. 446–473). Edward Elgar Publishing.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

Yang, V. C., Galesic, M., McGuinness, H., & Harutyunyan, A. (2021). Dynamical system model predicts when social learners impair collective performance. *Proceedings of the National Academy of Sciences*, *118*(35), e2106292118.