Preference-biased Social Influence*

Fabian Dvorak
University of Konstanz
fabian.dvorak@uni.kn

Urs Fischbacher
University of Konstanz
urs.fischbacher@uni.kn

November 22, 2021

Abstract

We propose a discrete-choice model that combines intrinsic preferences over choice alternatives with frequency-dependent social influence. The model assumes that the decision-maker has intrinsic preferences over a set of alternatives and observes the choices of a random sample of individuals from a reference group. Based on the observed choices, the decision-maker forms a belief about the frequencies of choices in the reference group and derives utility proportional to the natural logarithm of the expected frequencies. The model allows for variety of different reactions to social influence like conformity, non-conformity, independence or anticonformity and can be extended to accommodate situations in which the decision-maker's belief about the choice frequencies is biased by intrinsic preferences (motivated beliefs, false-consensus). We study the interplay of intrinsic preferences and social influence in an online experiment in which we measure participants' intrinsic preferences and provide information about others' behavior. We find that the model explains the average and individual behavior observed in the experiment well and substantially better than intrinsic preferences or social influence alone.

Keywords: conformity, social learning, preferences

JEL Classification: C92, D83, D91

*We would like to thank Brendan Barrett, Charles Efferson, the research group of the Thurgau Institute of Economics (TWI) and the Department of Economics at the University of Konstanz, as well as participants of various conferences, workshops and seminars for helpful comments and suggestions. All remaining errors are our own.

1 Introduction

A central question for models of human decision-making is whether the propensity to show a certain behavior varies with the prevalence of the same behavior in some peer or reference group. Such *social influence* can explain a variety of behavioral phenomena ranging from inter-dependent consumer demand (Gaertner, 1974; Pollak, 1976; Alessie and Kapteyn, 1991), over neighborhood effects (Schelling, 1971) and the emergence and persistence of social norms (Akerlof, 1980; Jones, 1984; Bernheim, 1994; Efferson et al., 2020) to behavioral nudges and peer effects (Allcott, 2011; Allcott and Rogers, 2014; Alpizar et al., 2008; Bobek et al., 2007; Coleman, 2007; Nolan et al., 2008; Schultz et al., 2007; Smith et al., 2015). A prominent formal model of social influence is frequency-dependent social influence (Boyd and Richerson, 1985) which has been successfully applied to explain behavior in humans and other animals (Aplin et al., 2017; Barrett et al., 2017; McElreath et al., 2005, 2008; Efferson et al., 2008; Toyokawa et al., 2019).

In this paper, we propose a model of human decision-making under social influence with intrinsic preferences. Individuals often have intrinsic preferences over choice alternatives that reliably predict individuals' choices in the absence of social influence (Smith et al., 2014). The interplay of intrinsic preferences and social influence defines which individuals are susceptible to social influence and affects the dynamics of social influence at the group-level. Additionally, humans often have reasonable prior beliefs about others' behavior and update their beliefs on a continuous basis when observing others' decisions. Using individuals' posterior beliefs as the basis of social influence can contribute to a better understanding of how individuals react to social information.

In the model, discrete choices depend on (1) intrinsic preferences over choice alternatives and (2) the *expected* frequency of each alternative in a reference group. The model is centered around two key assumptions. First, the model assumes that individuals derive intrinsic utility *and* social utility when making their decisions. This implies that individuals sometimes face a trade-off between what is intrinsically and socially preferred. Secondly, the model assumes that individuals have a prior belief about choices in the reference group and update this belief when observing a random sample of behavior. Social utility is a function of the individual's posterior belief about the frequency of choices in the reference group and *not* a function of the frequency of choices in the sample.

We study the interplay of intrinsic preferences and social influence in an online experiment. The exeriment extends the experimental design of Dvorak et al. (2020) and allows us to predict and manipulate participants' intrinsic preferences based on transitivity and to provide information about the behavior of five other participants without using deception. By varying the experimental task, we are able to explore whether individual-level responses to social influence differ between judgments on knowledge questions and preference-based selection of art paintings. We additionally conduct two between-subject treatments with social evaluation to manipulate participants' reactions to social information (Dvorak et al., 2020).

We find that the model explains the aggregate and individual data from the ex-

periment well and substantially better than netsted models which either take only intrinsic preferences or social influence into account. The parameter estimates suggest that both intrinsic preferences and social influence play an important role for participants' decisions. We find that social influence is task and treatment specific. Conformity prevails in both experimental tasks but is stronger in the experimental task in which participants' select answers to difficult questions with an objectively correct answer. We find that the experimental treatments with social evaluation affect indivdiuals' decisions through social influence. As in Dvorak et al. (2020) participants are generally more conformist in the treatment with punishment. In the treatment with reward, participants are generally less conformist. For the paintings task, the average reaction to social information shifts from conformity in the control treatment to non-conformity in the reward treatment.

Previous research on social influence often abstracts from intrinsic preferences and individuals' beliefs about others behavior. We argue that both factors are important to understand how social influence works. Manski (1993) highlights the fundamental problem of separating social influence from correlated effects, i.e. the fact that people behave similarly because they share similar individual characteristics. An important source of correlated effects are correlated intrinsic preferences over choice alternatives. If one alternative is superior to most people, it may look as if most individuals were conformists and imitate the majority - even if individuals' choices are independent. In such situations, controlling for individuals' intrinsic preferences is important to avoid wrong biased estimates of the effect of social influence.

The question whether social influence is based on observations or beliefs about others' behavior also plays a crucial role for the correct interpretation of the effect of social influence. For example, conformity to what is believed to be the majority behavior might be interpreted as anticonformity to sampled observations that are not representative for the individual's posterior belief. The assumption that social influence is based on beliefs about others' behavior can be plausible for several reasons. For example, individuals might have prior experience with others' behavior or might extrapolate from other experience. It may also be a natural assumption if the individual cares about the entire reference group and only observes random samples from the group or if others' decisions are subject to errors. In such situations, observations at a certain point in time can be 'surprising' as they are not in line with the individual's prior belief.[1]

Prior beliefs also relax the condition that small homogeneous samples elicit strong social influence. This is not only conceptually implausible but also econometrically challenging. When fitting frequency-dependent social influence models to the choices of individuals that are influenced by small samples, the models predict some

[1]The seminal studies of Asch (1952; 1956) on conformity provide a good example. One explanation why most participants did not conform to the wrong majority opinion is that it was in strong contrast to their prior expectation about the majority opinion. Aschs finding that the number of the majority increases the propensity to conform to the majority can also be explained by prior beliefs. Situations in which prior experience plays a crucial role are common when studying more "natural" choices outside the laboratory. Field studies on the effects of behavioral nudges provides a prime example. In these studies, researchers usually provide information about a sample of other individuals.

choices with certainty because observing small homogeneous samples generates exceptionally strong social influence. A single decision in the data that is not in line with the model's prediction will imply that the model has a likelihood of zero and makes it impossible to estimate parameters. Prior beliefs effectively reduce the impact of small samples and circumvent this problem.

Another advantage of explicitly modeling individuals' beliefs is that models of social influence can be extended to incorporate biases in the formation of beliefs (Bénabou and Tirole, 2002; Köszegi, 2006; Zimmermann, 2020). Such biases seem plausible since individuals frequently face a trade-off between intrinsic preferences and social influence and one way to resolve the conflict is to engage in self-deception. For example, a conformists could overestimate the frequency of her preferred alternative to resolve the conflict that arises when she intrinsically prefers the minority choice.

The model outlined in this paper is related to random utility models, and specifically, to the conditional logit model (Luce, 1959; McFadden, 1974; Revelt and Train, 1998). In the conditional logit model, discrete choices are a function of utility derived from observable characteristics of the choice alternatives. The preference part of the model assumes that choices generate alternative-specific utility proportional to the individual's intrinsic preferences. The social influence part of the model is a variant of frequency-dependent social influence (Boyd and Richerson, 1985). Frequency-dependent social influence allows for a continuum of individual-level responses to social information, which range from conformity over non-conformity and independence to anticonformity, and imply different choice dynamics at the group level, like polarized or mixed behavior (Efferson et al., 2008). An extension of frequency-dependent social influence is payoff-biased frequency-dependent social influence (McElreath et al., 2005; Baldini, 2012, 2013; Barrett et al., 2017), which assumes that social learning is frequency-dependent but biased by the observation of payoffs, received for a certain behavior. The social influence part of or model can be related to payoff-biased frequency-dependent social influence by assuming that individual learning has converged to time-invariant intrinsic preferences.

The paper is organized as follows. The following section introduces the model. Section 3 outlines the online experiments. Section 4 explains how we fit the model to the experimental data. The results are presented in Section 5. Section 6 concludes with a summary and discussion.

2 Model

Let Y_i denote the discrete choice of individual i among several alternatives with index $j \in \{1, \ldots, J\}$. Let U_{ij} represent the utility of alternative j to individual i. We assume that the utilities U_{ij} is the sum of two systematic components and one random component.

$$U_{ij} = \lambda_i v_{ij} + f_i \cdot log(s_{ij}) + \epsilon_{ij} \tag{1}$$

The two systematic components are the elements $\lambda_i v_{ij}$, and $f_i \cdot log(s_{ij})$ of the function above. The component $\lambda_i v_{ij}$ reflects the intrinsic utility of alternative j for the individual. As in the conditional logit model (McFadden, 1974), it is assumed that intrinsic utility is the product of several observable characteristics and their marginal utilities. To simplify the notation, we consider a single characteristic v_{ij} multiplied by λ_i, the marginal utility of this characteristic. To give an example, if alternative j is an object, potential observable characteristics could be the shape, form or functionality of the object.

The second systematic component $f_i \cdot log(s_{ij})$ reflects the social utility of the alternative. Social utility is the individual-specific parameter f_i, that reflects individual i's (anti)conformity type, multiplied by the logarithm of the expected share s_{ij} of alternative j in the reference group.

The component ϵ_{ij} captures randomness in the preferences of individual i for the jth alternative. We will assume that ϵ_{ij} is distributed iid extreme value I (Gumbel) which results in choice probabilities in the logit form (Luce, 1959; McFadden, 1974). If we assume $\epsilon_{ij} \sim Gumbel(0, 1)$, the probability that individual i chooses alternative j has the following closed-form solution:

$$Pr(Y_i = j) = \frac{e^{\lambda_i v_{ij}} (s_{ij})^{f_i}}{\sum_k e^{\lambda_i v_{ik}} (s_{ik})^{f_i}} \tag{2}$$

Equation (2) shows that the systematic components of the utility function influence the choice probabilities based on different functional forms. For the characteristics v_{ij}, the functional form is exponential. The exponential form has the advantage that the characteristics v_{ij} can be specified in absolute or relative terms.

Using the logarithm of the expected shares in the utility function (1) yields the power form for the expected shares. A practical advantage of the power form is that is invariant to multiplying all shares by a constant. As a result, the model makes the same predictions if one uses expected counts instead of expected shares - which is conceptually plausible. Additionally, the power form allows for a variety of functional forms over s_{ij} which includes sigmoid, linear and inverse sigmoid and other shapes. Each functional form implies different choice dynamics in a system of repeatedly interacting agents as discussed in the literature on frequency-dependent social influence (see Efferson et al., 2008, for example).

The model defined by Equation (1) nests two prominent models of behavior. The conditional logit model is obtained for $f_i := 0$. The frequency-dependent social influence model is obtained for $\lambda_i := 0$ together with the assumption that the expected shares s_{ij} correspond to the frequencies of choices observed in the sample.

2.1 Beliefs about shares

As the share of individuals in the reference group that chooses alternative j is unknown, individual i holds a prior belief about the share of alternative j. We assume that the prior belief follows a Dirichlet distribution with parameters $\alpha = \alpha_{i1}, \ldots, \alpha_{iJ}$. The α-parameters define the shape of the prior belief distribution and

the expected shares $s_{ij} = \alpha_{ij} / \sum \alpha_{ik}$. If $\alpha_{ij} = 1 \; \forall j \in \{1, \ldots, J\}$, the prior belief is uniform, all shares in the unit interval are believed to be equally plausible. In this case, the expected share is $1/J$. The larger α_{ij} is relative to the other α-parameters, the more likely are large shares of alternative j.

After observing a random sample of N individuals of which $n_{ij} \in \{0, \ldots, N\}$ prefer alternative j, individual i updates the prior belief about the share of alternative j in the reference group. Since the Dirichlet distribution is the conjugate prior of the multinomial distribution, the resulting posterior follows a Dirichlet distribution with updated parameters $\alpha_{ij} + n_{ij}$. After the belief update, the expected shares of the alternatives are

$$s_{ij} = \frac{\alpha_{ij} + n_{ij}}{\sum_k^J (\alpha_{ik} + n_{ik})}. \tag{3}$$

The utility defined in Equation (1) differs from the expected utility of the alternative given the posterior belief of the individual. Equation (1) suggests that the individual does *not* form an expectation about the social utility of the alternatives. Instead, the individual first updates the prior belief when exposed to social information, calculates the expected share of each alternative in the reference group, and then decides based on the utilities of the alternatives that result from the expected shares.[2]

2.1.1 Biased beliefs

An advantage of explicitly modeling individuals' beliefs is that the model can be extended to account for biases in the belief-formation process. From Equation (1) it is clear that an individual might be in a situation in which her intrinsic preferences and social influence suggest different alternatives. For example, consider a conformist individual who strongly prefers one alternative but, at the same time, believes that the other alternative is more frequent in the reference group. Depending on the strength of her intrinsic preferences and the strength of social influence, the individual will either stick to the intrinsically preferred alternative at the cost of less social utility or switch to the more popular alternative at the cost of less

[2]Expected utility maximization requires that the individual maximizes $E(log(X))$ where X is a Dirichlet distributed random variable with parameters $\alpha_{ij} + n_{ij}$ and:

$$E(log(X)) = \psi(\alpha_{ij} + n_{ij}) - \psi\left(\sum_k \alpha_{ik} + n_{ik}\right)$$

where $\psi(\cdot)$ is the digamma function which has the logistic approximation $\psi(x) \approx log\left(x - \frac{1}{2}\right)$ (Johnson et al., 1994). Using this approximation:

$$E(log(X)) \approx log\left(\frac{\alpha_{ij} + n_{ij} - \frac{1}{2}}{\sum_k (\alpha_{ik} + n_{ik}) - \frac{1}{2}}\right) \approx s_{ij}$$

if the quantity $\alpha_{ij} + n_{ij}$ is large for all alternatives.

intrinsic utility. One way to resolve such conflicts is by manipulating beliefs in a way that the social utility penalty for the intrinsically preferred alternative is less severe. Such act of self-deception are known as false-consensus effect (Mullen et al., 1985) or motivated beliefs (Bénabou and Tirole, 2006).

One possibility to incorporate motivated beliefs is to define the parameters of the prior belief distribution as:

$$\alpha_{ij} = \phi_i \cdot J \cdot \frac{e^{\delta_i v_{ij}}}{\sum_k e^{\delta_i v_{ik}}} \tag{4}$$

where $\phi_i > 0$ is the strength of prior belief and δ_i is a parameter for the preference bias. If the parameters δ_i and λ_i have the same sign, the prior is biased such that large shares of the intrinsically preferred alternative are more likely. If the signs of the two parameters δ_i and λ_i differ, the prior is biased such that small shares of the intrinsically preferred alternative are more likely. If δ_i is zero, the individual's prior belief is unbiased and the expected shares are $1/J$ for every alternative.

Other variants of preference-biased prior beliefs are also plausible.[3] We use the form outlined in Equation (4) because of the convenient property that the prior expected share is:

$$s_{ij} = \frac{\alpha_{ij}}{\sum_k \alpha_{ik}} = \frac{e^{\delta_i v_{ij}}}{\sum_k e^{\delta_i v_{ik}}}$$

If $\delta_i \approx \lambda_i$, this means that the expected prior belief corresponds to the own behavior in the absence of social influence.

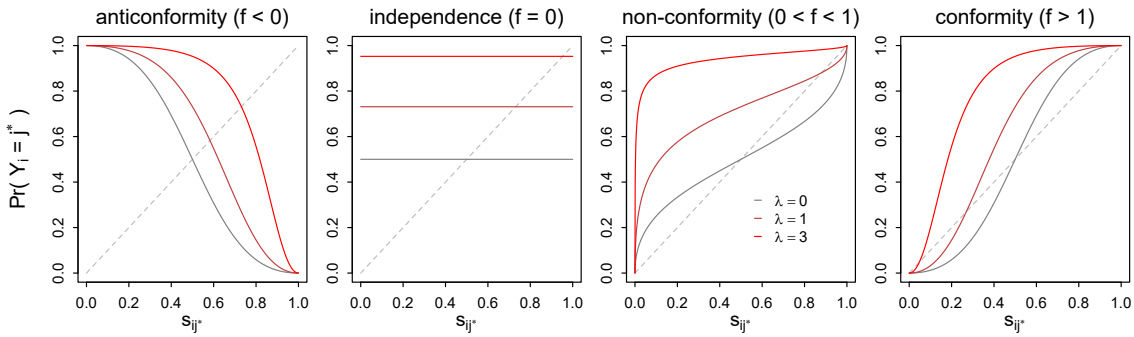
2.2 Comparative statics

The model conveys several stylized facts about social influence. The four graphs of Figure 1 illustrate the effects of the parameters λ and f on the probability to choose the preferred alternative. We use the index j^* to indicate the preferred alternative. The probability to choose alternative j^* is a function of the expected share of alternative s_{j^*} in the reference group depicted on the x-axis. In Figure 1, this function is plotted for a binary choice. Note that we omit the individual index i for better readability.

The four different panels illustrate the shape of the probability function for different values of the social influence parameter f. Conditional on the value of the social influence parameter, the individual is classified as anticonformist, independent, non-conformist or conformist. While anticonformity (left panel) implies that the probability to choose the preferred alternative decreases in the expected frequency of the preferred alternative in the reference group, non-conformity and conformity

[3] An alternative way to introduce preference-biased beliefs would be to assume that the updating process of the individual is biased by her intrinsic preferences (Zimmermann, 2020). This can be achieved by introducing updating weights proportional to the intrinsic utility of each alternative.

Figure 1: Predicted response to social influence



Notes: The four graphs show the probability of choosing the intrinsically preferred alternative j^* in a binary choice as a function of the expected frequency of the preferred alternative j^* in the reference group (s_{j^*}). From left to right, the graphs show this probability for an anticonformist, independent, non-conformist and conformist individual, characterized by the social influence parameter (f). The solid lines assume that v_j is a dummy for the preferred alternative and vary the λ parameter.

(central-right and right panels) both imply a positive relation between the probability to choose alternative j^* and the expected share of j^* in the reference group. Independence (central-left panel) implies that the probability to choose alternative j^* is not affected by the expected frequency of the alternative in the reference group.

The solid red lines in Figure 1 show how the probability to choose alternative j^* is influenced by the model parameter λ. The more red the color of a line is, the stronger is the utility difference between the alternatives. Figure 1 illustrates that intrinsic preferences bend the probability curves upwards which makes it more likely that the individual chooses the intrinsically preferred alternative.

Figure 1 illustrates three stylized facts about social influence:

1. *Intrinsic preferences immunize against social information.* Figure 1 shows how intrinsic preferences mediate the response to social information. The key aspect to observe is that intrinsic preferences for alternative j^* (solid red lines) create a larger region of possible beliefs for which the predicted response curve is relatively high and flat, i.e. insensitive to changes in the belief. This implies that (1) the individual almost certainly chooses the intrinsically preferred alternative j^* and (2) needs a relatively large and homogeneous sample to overcome this tendency.

2. *Strong prior beliefs immunize against social information.* Figure 1 also illustrates the role of prior beliefs for individuals choices. Before observing a sample of others' behavior, the choice probability is determined the expected share of the preferred alternative in the reference group. When observing the sample, the individual updates the prior belief. If the observed frequencies differ from the individual's prior belief, the expected share will also differ and the choice probability will change accordingly. The strength of the individual's prior belief defines the distance between the prior and the posterior expected share on the x-axis of Figure 1. If choices are not independent ($f \neq 0$), a strong prior implies a small change in the choice probability before and after

observing the sample. Conversely, this means that, given some strength of the prior belief ($\phi > 0$), the effect of social information increases in the size of the sample.

3. *Majority behavior contains information about intrinsic preferences.* Each response curve depicted in Figure 1 implies different choice dynamics in a group of repeatedly interacting individuals (see Efferson et al., 2008, for details). The intersections of the solid lines with the dashed lines fix points of the choice dynamics, assuming that all individuals are the same and individuals' beliefs will eventually converge towards the true choice frequencies. For $f < 1$ the red lines suggest converges to a state in which the majority of individuals chooses the intrinsically preferred alternative. In this state of convergence, the frequency of the majority behavior is indicative of the strength of the intrinsic preference for the majority choice. For conformity ($f > 1$) the strength of the intrinsic preference increases the basin of attraction of the majority choice which means that the group converges to the intrinsically preferred choice with higher probability. Independent of the type of social influence, the behavior of the group contains information about intrinsic preferences of the group after convergence.

3 Experiment

We study the interplay of intrinsic preferences and social influence in an online experiment. We use the experimental design of Dvorak et al. (2020) that allows to infer participants' intrinsic preferences and provide information about others' behavior without using deception. In the original design, participants receive information about the choices of two other participants of the experiment. We extend the original design such that participants can be informed about the choices of five other participants. This makes it possible to model participants' choices as a function of the number of other participants that choose the same alternative.

3.1 Choice tasks

In the online experiment, participants made binary choices in two different choice tasks. The first choice task is to select one of two artistic paintings from well-known artists. The second experimental task is to select the correct answer of two possible answers to a knowledge question. A list of all paintings, questions and answers we used can be found in the Appendix of Dvorak et al. (2020).

Each participant of the experiment made 60 binary decisions in each experimental task. 50 decisions are made without information about the choices of other participant. 10 decisions are made with information about the choices of other participants. We use the 50 decisions to measure the intrinsic preferences of the participant in the 10 decisions with social information. To incentivize participants' decisions in the painting selection task, participants received the motive selected in

one of the binary choices printed on a postcard. The postcard was sent to the participant in an envelope several days after the experiment. To increase the intrinsic motivation in the questions task, participants were informed about the objectively correct answers at the end of the experiment.

3.2 Measurement of intrinsic preferences

The left panel of Figure 2 illustrates how we infer participants' intrinsic preferences in the decisions with social information. There are always four choice alternatives A, B, C, D. To measure the intrinsic preference Δ of a participant in the binary choice between alternatives A and B, we use the pairwise comparisons of both alternatives to the two other alternatives C and D. These yield the four *related* binary choices A vs. C, B vs. C, A vs. D and B vs. D which are depicted in the left panel of Figure 2. After each of the four choices, the participant indicates the strength of her preference using a slider ranging from minus one (strong preference for the alternative displayed on the left), over zero (indifference) to one (strong preference for the alternative displayed on the right). The slider appeared at the bottom of the decision screen after the decision. An example of the decision screen with the slider can be found in the Appendix.

The self-reported preference strength in the two related choices that compare the alternatives A and B to a common third alternative yield an estimate for the participant's intrinsic preferences in the decision A vs. B. Since there are two comparisons of A and B to a common third alternative, we use the expected value of the two estimates as a measure for the participant's intrinsic preferences in the decision A vs. B.

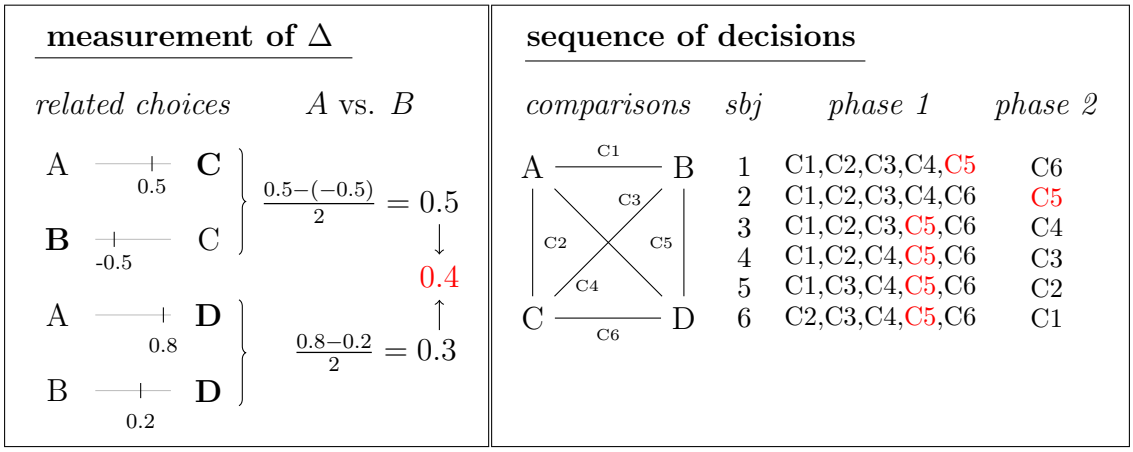
In the example in the left panel of Figure 2, the participant chooses alternative C over A and reports a preference strength of 0.5. In another related binary choice, the same participant preferred B over C and reported a preference strength of -0.5. The difference between the two reported values divided by two produces an estimate for the intrinsic preference of the participant in the choice A vs. B. The division makes sure that the estimate is between minus one and one. The two other related binary comparisons A vs. D and B vs. D yield an estimate of 0.3. In the example, we would use the expected value $\frac{0.5+0.3}{2}$ of the two estimates as a measure for the intrinsic preference of the participant in the decision A vs. B.

The sign of Δ indicates the intrinsically preferred alternative of the participant. In the example, the sign of the expected value is positive which indicates that alternative B is intrinsically preferred. The absolute difference in intrinsic utility between the two alternatives is measured by Δ, the absolute value of the expected value. The variable Δ ranges between zero and one. In the example, the estimated absolute difference in intrinsic utility between alternatives A and B is 0.4 - which is less than half of the maximum preference strength.

The method to infer intrinsic preferences illustrated in Figure 2 rests on the assumption of transitivity across the four choice alternatives. The method has the advantage that the participants of the experiment do not have to make the same

binary choice twice - once without and once with social influence. This reduces the concerns that participants want to make consistent choices, which would lead to an underestimation of the effect of social influence. The technique to measure intrinsic preferences presented in the left panel of Figure 2 produces estimates that are proportional to the true intrinsic preferences plus some independent, identically distributed random error as long there is no systematic violation of transitivity.

Figure 2: Experimental design



Notes: Left panel illustrates the method to measure the intrinsic preference a participant in the binary decision A vs. B. The estimate Δ of the intrinsic preference is the expected value of two normalized estimates that are derived by comparing the two alternatives A and B to a common third alternative - either C or D. The division by two normalizes the two estimates such that Δ ranges from -1 to 1. Right panel illustrates how the six pairwise comparisons of four alternatives are distributed over two phases of the experiment. In phase 2, each participant is informed about the uninfluenced choices of the five other participants in phase 1.

3.3 Social information

Before a participant makes the the decision with social influence between A and B, the participant receives information about the decisions of five other participants in the same choice. Depending the experimental task, participants see the selected paintings or the selected answers of the five other participants on the decision screen. An example of the decision screen can be found in the Appendix.

At the time of their decisions, the five other participants themselves have no information about the decisions of other participants. The right panel of Figure 2 illustrates how the experimental design efficiently makes sure that this condition is fulfilled for a group of six subjects. The pairwise comparisons of four alternatives A, B, C and D generate six binary decisions. In a first phase of the experiment, each participant of a group of six completes five of the six binary decisions without social information. In phase two of the experiment, the remaining sixth decision is completed by each participant. Since this sixth decisions differs for each participant

of the group, it is possible to inform the participant about the choices of the other five participants from phase one.

In the example in the right panel of Figure 2, the decisions highlighted in red mark the decision C5 which is the comparison of alternatives B and D. All subjects except subject 2 complete this decision in phase 1. Subject 2 can therefore be informed about the decisions of the five other subjects before she makes her decision in phase 2. The right panel of Figure 2 illustrates that this works for all six subjects.

3.4 Experimental treatments

As in Dvorak et al. (2020), we conduct two experimental treatments with social evaluation and a control treatment without social evaluation. In the treatments with social evaluation, participants are informed that their decisions with social information, together with the choices of the five other participants, will be shown to a participant who not part of the group. This participant takes the role of an evaluator. The six choices are displayed on the screen of the evaluator in a randomized order. The task of the evaluator, who does not know which of the displayed choices is the social choice, is to select one of the six participants by clicking on one of the six choices. At the end of the experiment, one decision of the evaluator is randomly chosen for each experimental task. In the reward treatment, the payoff of the participant who was selected in the randomly chosen evaluator decision is increased by 10 Euro. In the punishment treatment, the payoff of the selected participant is reduced by 10 Euro. An example of the evaluation screen can be found in the Appendix.

Dvorak et al. (2020) study a simple model of social evaluation in which the evaluator selects a participant based on her own intrinsic preference. The evaluator selects a participant who has chosen the preferred alternative of the evaluator for reward and a participant who has *not* chosen the preferred alternative for punishment. With information about others behavior, the incentive to choose alternative j depends the number n_{ij} of other participants that have chosen alternative j in phase 1 of the experiment. In the reward treatment, the incentive to choose alternative j generally decreases in n_{ij} - which incentivizes anti-conformity. In the punishment treatment, the incentive to choose alternative j increases in n_{ij}, which incentivizes conformity. The intuition is that, since the social evaluation is based on participants' choices, reward and punishment is shared among those who display the same behavior. Therefore, we expect the treatments with social evaluation to influence the model parameter f that captures the reaction of a participant to n_{ij}, the number of other individuals that have chosen alternative j. Compared to the control treatment, the social influence parameter f should be larger under punishment and smaller under reward.

3.5 Implementation

We conducted 10 experimental sessions online between December 11-16, 2020 with 745 students of various fields of the University of Konstanz in Germany. The mean age of the participants was 21.6 years. We recruited the participants of the experiment with the recruitment software *hroot* (Bock et al., 2014). The experiments were conducted with *z-Tree* (Fischbacher, 2007) and *z-Tree unleashed* (Duch et al., 2020). The duration of the experimental sessions was between 1.5 and 2 hours. The data set contains 3475 binary choices under social influence. Participants received a fixed amount of experimental currency for each decision in the experiment. On average, a participant earned 23 Euro for participating in the experiment.

4 Statistical model

We fit a Bayesian multilevel model to the data of the online experiment using *R* (R Core Team, 2021), *Stan* (Team, 2021) and the packages *RStan* (Stan Development Team, 2020) and *rethinking* (McElreath, 2020). The multilevel model includes fixed effects for the experimental task (index t) and individual random effects (index i) for all model parameters. We use and additional fixed effect for the treatment condition (index c) for the social influence parameter f to capture the effect of the experimental treatment on this model parameter. We use standard priors that result in expected effects of zero which corresponds to random choice and Hamiltonian Monte-Carlo (Neal, 2011; Betancourt, 2013) to approximate the posterior distributions of the parameters.[4]

We implement the model proposed in Section 2 with preference biased prior beliefs. In the statistical model, the probability that participant i chooses alternative j is:

$$Pr(Y_i = j) = \frac{e^{\lambda_{it}\nu_{ij}}(s_{ij})^{f_{itc}}}{\sum_k e^{\lambda_{it}\nu_{ik}}(s_{ik})^{f_{itc}}}$$

The model parameter λ_i reflects the effect of the estimated intrinsic utility ν_{ij} on the individual's choice. The exponential form implies that the choice probability is a function of the difference in intrinsic utility of both alternatives that we measure by Δ. For convenience we use $\nu_{ij} := \frac{\Delta}{2}$ if j is predicted to be the preferred alternative, and $\nu_{ij} := -\frac{\Delta}{2}$ otherwise. The exponential form of the intrinsic preference

[4] We use the standard normal distribution for the prior of the fixed effects of all model parameters except ϕ which cannot be negative. For the parameter ϕ, we use $exp(N(0,1))$. As prior for the individual varying effects, we use the multivariate normal distribution $MVN(0,\Sigma)$ with:

$$\Sigma = \begin{pmatrix} \sigma_\lambda & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_\delta \end{pmatrix} \Omega \begin{pmatrix} \sigma_\lambda & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_\delta \end{pmatrix}$$

where $(\sigma_\lambda, \sigma_f, \sigma_\phi, \sigma_\delta) \sim Exponential(1)$ and $\Omega \sim LKJcorr(2)$.

component assures that different distributions of Δ over the two choice alternatives will give the same results.

The model parameter f_{itc} reflects the effect of social influence based on s_{ij}, the frequency of alternative j expected by the participant. The expected frequency of alternative j is:

$$s_{ij} = \frac{2\phi_{it}\frac{e^{\delta_{it}\nu_{ij}}}{\sum_k e^{\delta_{it}\nu_{ik}}} + n_{ij}}{2\phi_{it} + 5}$$

The parameter ϕ_{it} is the strength of the prior and can be interpreted as the number of past observations of each alternative. The parameter δ reflects the preference bias of the prior belief. As before, n_{ij} indicates the number of observed uninfluenced participants that chooses alternative j.

5 Results

5.1 Model fit

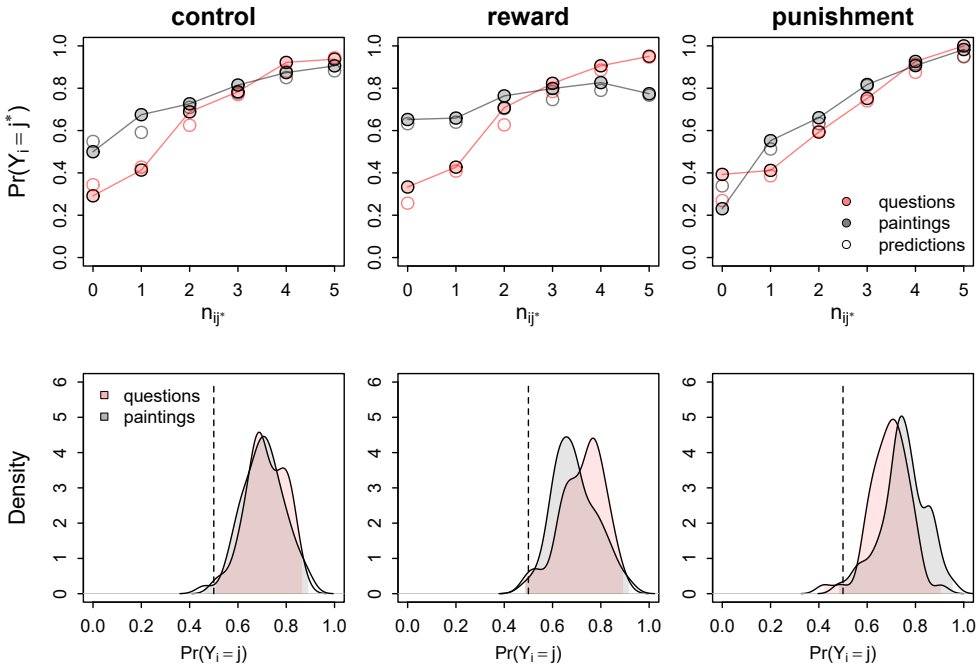
Figure 3 illustrates the model fit for average and individual behavior. The three plots in the first row of Figure 3 plot model predictions (circles) against observed average behavior (dots) for the two experimental tasks in each of the three experimental treatments. The filled dots depict the average probability that a participant chooses her intrinsically preferred alternative j^* conditional on n_{ij^*} - the number of other participants that have chosen this alternative. The model predictions are fairly close to the observed average behavior in the two experimental tasks of each treatment. The fit is generally better for larger values of n_{ij^*}. There seems to be no systematic error in the model predictions.

To illustrate the fit of the model to the individual behavior, we consider all choices from the same individual in the same task and calculate the mean posterior probability of these choices for different sets of model parameters sampled from the posterior distributions. The three graphs in the lower part of Figure 3 depict the distribution of the mean posterior probability of participants' choices conditional on the experimental task and treatment. In all three graphs, most of the density mass of is above one-half - which is the expected benchmark for random predictions indicated by the dashed line. Only few individuals fall below this benchmark. The mode of the mean posterior probability is around 70%.

5.2 Parameter estimates

Table 1 shows estimated posterior means of the model parameters along with the 5th and 95th percentile of the posterior distribution. For the social influence parameter f, the table shows the estimates for the control treatment without social evaluation. The posterior means of all model parameters are larger than zero which suggests that participants' decisions are influenced by intrinsic preferences, social influence

Figure 3: Fit for average and individual behavior



Notes: Upper part shows model fit to average behavior. Circles represent model predictions. Dots represent average behavior observed conditional on the information about others behavior. Lower part shows the distribution of the mean posterior probability of all the choices from the same participant.

and preference-biased prior beliefs. The size of the parameter estimates varies between the experimental tasks.

Table 1: Posterior distributions of model parameters

	λ		f		ϕ		δ	
task	quest	paint	quest	paint	quest	paint	quest	paint
mean	2.89	5.09	2.61	1.22	3.29	1.98	2.89	2.36
5th percentile	0.99	3.25	1.81	0.57	1.66	0.48	1.47	0.32
95th percentile	4.72	6.75	3.55	2.13	5.72	5.93	4.36	4.18

Notes: Table shows means and percentiles of the posterior distribution of the model parameters. The values for the social influence parameter f are reported for the control treatment.

The posterior means of the parameter λ indicate that the measured intrinsic preferences influence participants' decisions. All estimates are clearly positive which indicates a tendency to choose the alternative with greater intrinsic utility. The posterior means of the social influence parameter f indicate conformity in both experimental tasks for the control treatment. The estimates suggest more conformity in the questions task. The estimates of the belief parameters ϕ and δ suggest that participants hold preference-biased prior beliefs. The estimates of δ are positive. On average, participants think that it is more likely that their intrinsically preferred alternative is chosen by more individuals of the reference group.

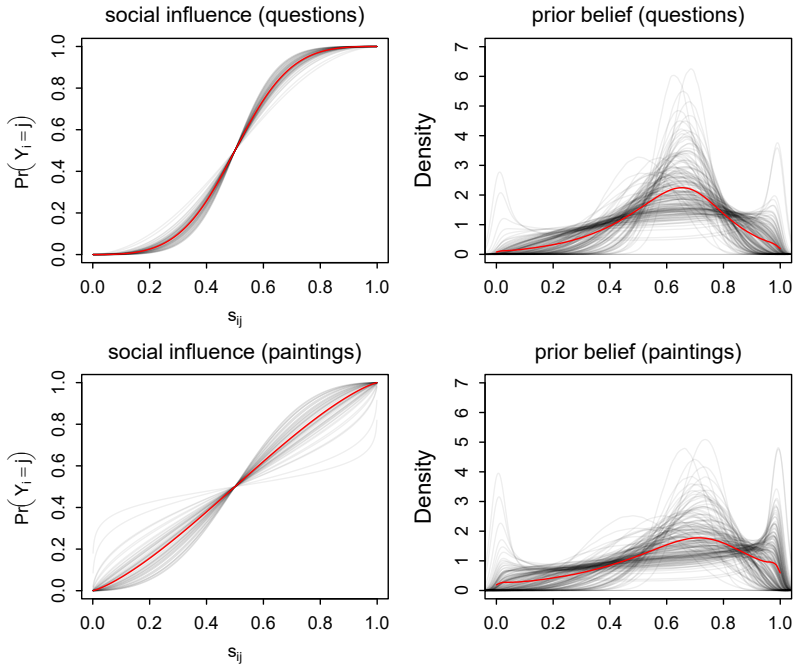
5.3 Individual heterogeneity

Figure 4 depicts individual heterogeneity in social influence (left graphs) and prior beliefs (right graphs) for the two experimental tasks. Each line represents an individual estimate. The red curves indicate the mean of the individual curves.

The upper left graphs of Figure 4 suggests that participants' conformity prevails in the questions task with little individual heterogeneity. The inter-individual differences are more pronounced in the paintings task. In this task, the majority of individuals is conformist and displays an S-shaped reaction to the expected frequency of an alternative. A minority is non-conformist with an inverse S-shaped reaction.

The right graphs depict individuals' prior beliefs about the frequency of their intrinsically preferred alternative in the reference group. The graphs suggest that majority of individuals discount the information presented to them about the frequency of choices. Some individuals hold a U-shaped prior which suggests no discounting of social information. The depicted prior beliefs also show that most individuals hold the preference-biased prior belief that larger frequencies are more plausible for their intrinsically preferred alternative.

Figure 4: Individual parameter estimates

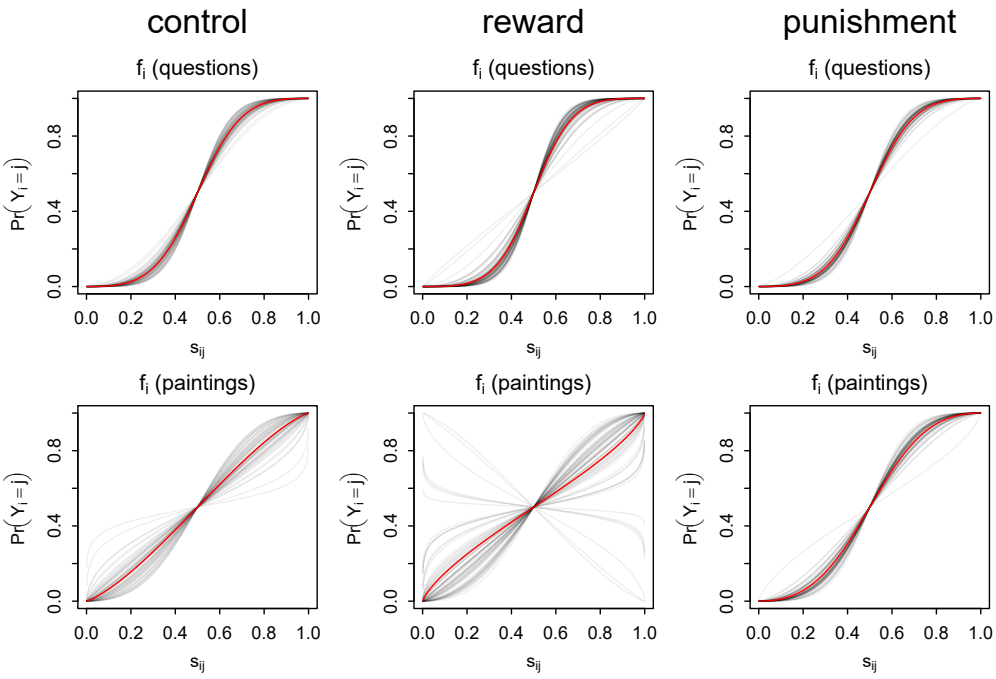


Notes: Figure depicts individual estimates for social influence (left graphs) and prior beliefs (right graphs). Red lines indicate the average of the individual estimates.

5.4 Treatment effects

As in Dvorak et al. (2020), we find that the evaluators allocate punishment and reward based on their own intrinsic preferences. Based on their theory and experimental findings, we expect that punishment induces conformity while reward induces anticonformity. This means that, compared to the control treatment, the social influence parameter f should be larger under punishment and smaller under reward.

Figure 5: Treatment effect on social influence



Notes: Figure depicts individual estimates for social influence for the two experimental tasks (rows) in each treatment condition (columns). Red lines indicate the average of the individual lines.

Figure 5 shows that this is the case for the paintings task. The posterior mean increases from X in the control treatment (left) to Y under punishment (right). This effect is associated with a reduction of inter-individual differences in the social influence parameter f. In the control treatment, some individual estimates indicate conformity $f > 1$ and others indicate non-conformity $f < 1$, the frequency of non-conformists is reduced. In the reward treatment (center), the average behavior is non-conformist and more individual heterogeneity arises. A minority of individual estimates indicate anticonformity $f < 0$, which is in line with previous findings Dvorak et al. (2020).

For the questions task, the treatments effects are not consistent with the expectations despite some individual heterogeneity under reward which goes into the right direction.

5.5 Model comparisons

We compare the model to three nested models. The first model does not include a parameter for the preference-bias (PSI model). Instead, it uses $s_{ij} = (\phi_{it} + n_{ij})/(2\phi_{it} + 5)$. The second nested model is a pure preference model that only takes the intrinsic preferences over the choice alternatives into account (P model). The third model is a pure social influence model in which participants' choices depend on the expected frequencies of the choice alternatives (SI model). The three nested models are obtained from the preference-biased social influence model (PBSI model) outlined in Section 4 for the restrictions $\delta := 0$, $f := 0$ and $\delta := 0 \wedge \lambda := 0$ respectively.

Table 2 summarizes the posterior means of the model parameters for all four models. The posterior means of the preference parameter λ are larger in the PSI and P models and do not differ between the experimental tasks. The posterior means of the social influence parameters f estimated for the control treatment are smaller in the PSI and SI models and suggest a non-conformist reaction to social information in the paintings task. The estimates of the parameter ϕ - which indicates prior experience - are larger in these models.

Table 2: Posterior means

	λ		f		ϕ		δ		WAIC
task	quest	paint	quest	paint	quest	paint	quest	paint	
PBSI	2.89	5.09	2.61	1.22	3.29	1.98	2.89	2.36	3157
PSI	6.28	6.81	2.08	0.58	1.55	0.29	-	-	3180
P	6.36	6.68	-	-	-	-	-	-	3652
SI	-	-	1.94	0.69	1.70	0.39	-	-	4192

Notes: Table compares posterior means of the model parameters of four different models. The last column depicts the value of the widely applicable information criterion (Watanabe, 2010) for each model.

The last column of Table 2 indicates the values of the widely applicable information criterion (WAIC, Watanabe, 2010) for the four models.[5] The WAIC approximates the out-of-sample prediction accuracy of a Bayesian model. It is asymptotically equivalent to leave-one-out cross-validation (Vehtari et al., 2017). Hence, models with low WAIC values are preferable. The WAIC values indicate that the model with intrinsic preferences, social influence and preference-biased prior beliefs generates more accurate out-of-sample predictions than the nested models.

[5] The formula of the WAIC is:

$$WAIC(Y|\Psi) = -2\left(\sum_m log\frac{1}{S}\sum_s Pr(Y_m|\Psi_s) - \sum_m var\left(log\left(Pr(Y_m|\Psi_s)\right)\right)\right)$$

where Y are the observed choices and Ψ are the samples used to approximate the posterior (McElreath, 2020).

We also fit models with fixed effects of the treatment condition for all model parameters. We find that these models do not yield better out of sample predictions according to the WAIC. This means that the treatment effect is best explained by the social influence parameter. Removing the task fixed effect of any model parameter also results in larger values of the WAIC.

6 Summary and Discussion

We introduce a discrete-choice model that combines intrinsic preferences over choice alternatives with frequency-dependent social influence. The model addresses two limitations of models of social influence. First, models of social influence can produce wrong estimates of the nature of social influences in situations in which individuals have intrinsic preferences. Secondly, inferences from these models can also be flawed if individuals have prior beliefs about the choice frequencies in a reference group. We propose a model of decision-making which is useful in such situations and can also be applied when individuals' beliefs about others' behavior are biased by their intrinsic preferences. As a convenient feature, the model nests conditional logistic choice and frequency-dependent social influence and can therefore be applied to test for the influence of intrinsic preferences or preference-biased prior beliefs.

We demonstrate the usefulness of the model by fitting it data from an online experiment. The parameter estimates indicate that intrinsic preferences and social influence both influence participants' decisions in the experiment. We show that the model fits the average and individual behavior observed in the experiment well and performs better than three nested models that do not account for biased beliefs, participants' intrinsic preferences or social influence respectively.

The model can explain several stylized facts about social influence. The first concerns the questions under which circumstances people are *not* susceptible to information about others' behavior. The model suggests that those who are *not* prone to social information either have strong intrinsic preferences (idealists) or substantial prior knowledge about others' behavior (strong believers). Idealists are somewhat immune to social information as their intrinsic preference is too strong that social influence could overrule it. Strong believers on the other hand do not react to social information since their prior belief is too strong to be substantially affected by information about others' behavior.

Since the model accounts for prior experience, it can explain why the propensity for conformity increases in the size of the majority - a finding which pervades the experimental literature on conformity (Bond and Smith, 1996; Sasaki, 2019). The reason is that, given the same prior belief, large samples have a greater impact on the individual's posterior belief. Prior experience can also explain social inertia in behavior which is sometimes observed when social influence is supposed to induce social change. If the status-quo exist for a long time, individuals develop strong beliefs that, in combination with conformity, stabilize the status-quo (Smerdon et al., 2020). The combination of intrinsic preferences and prior experience suggests that idealistic individuals with different intrinsic preferences initiate and young

individuals with weak prior beliefs drive social change.

Finally, the model also explains why it can be reasonable to infer intrinsic preferences from average behavior even if individuals have repeatedly interacted in the past and socially influence each other. For example, if all individuals are conformists with the same intrinsic preferences, it is most likely that the group's choices converge to the intrinsically preferred behavior. If the individuals are non-conformists, choices will be mixed but the intrinsically preferred behavior will be more frequent. This provides a justification for the viewpoint that social norms are not arbitrary but have some intrinsic value which goes beyond the benefits of coordination.

References

Akerlof, G. A. (1980): "A Theory of Social Custom, of which Unemployment may be One Consequence*," *The Quarterly Journal of Economics*, 94, 749–775.

Alessie, R. and A. Kapteyn (1991): "Habit Formation, Interdependent Preferences and Demographic Effects in the Almost Ideal Demand System," *The Economic Journal*, 101, 404–419.

Allcott, H. (2011): "Social norms and energy conservation," *Journal of Public Economics*, 95, 1082 – 1095.

Allcott, H. and T. Rogers (2014): "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation," *American Economic Review*, 104, 3003–37.

Alpizar, F., F. Carlsson, and O. Johansson-Stenman (2008): "Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica," *Journal of Public Economics*, 92, 1047 – 1060.

Aplin, L. M., B. C. Sheldon, and R. McElreath (2017): "Conformity does not perpetuate suboptimal traditions in a wild population of songbirds," *Proceedings of the National Academy of Sciences*, 114, 7830–7837.

Asch, S. (1952): *Social Psychology*, New-Jersey: Prentice-Hall.

Asch, S. E. (1956): "Studies of independence and conformity: A minority of one against a unanimous majority," *Psychological Monographs*, 70.

Baldini, R. (2012): "Success-biased social learning: Cultural and evolutionary dynamics," *Theoretical Population Biology*, 82, 222–228.

——— (2013): "Two success-biased social learning strategies," *Theoretical Population Biology*, 86, 43–49.

Barrett, B. J., R. L. McElreath, and S. E. Perry (2017): "Pay-off-biased social learning underlies the diffusion of novel extractive foraging traditions in a wild primate," *Proceedings of the Royal Society B: Biological Sciences*, 284, 20170358.

Bénabou, R. and J. Tirole (2002): "Self-Confidence and Personal Motivation*," *The Quarterly Journal of Economics*, 117, 871–915.

Bénabou, R. and J. Tirole (2006): "Incentives and prosocial behavior," *American Economic Review*, 96, 1652–1678.

Bernheim, B. D. (1994): "A Theory of Conformity," *Journal of Political Economy*, 102, 841–877.

Betancourt, M. (2013): "A General Metric for Riemannian Manifold Hamiltonian Monte Carlo," in *Geometric Science of Information*, ed. by F. Nielsen and F. Barbaresco, Berlin, Heidelberg: Springer Berlin Heidelberg, 327–334.

Bobek, D., R. Roberts, and J. Sweeney (2007): "The social norms of tax compliance: evidence from Australia, Singapore and the United States," *Journal of Business Ethics*, 74, 49–64.

Bock, O., I. Baetge, and A. Nicklisch (2014): "hroot: Hamburg Registration and Organization Online Tool," *European Economic Review*, 71, 117–120.

Bond, M. H. and P. B. Smith (1996): "Cross-Cultural Social and Organizational Psychology," *Annual Review of Psychology*, 47, 205–235, pMID: 15012481.

Boyd, R. and P. J. Richerson (1985): *Culture and the evolutionary process*, Chicago: University of Chicago Press.

Coleman, S. (2007): "The Minnesota Income Tax Compliance Experiment: Replication of the Social Norms Experiment," Tech. rep., Available at SSRN: https://ssrn.com/abstract=1393292 or http://dx.doi.org/10.2139/ssrn.1393292.

Duch, M. L., M. R. Grossmann, and T. Lauer (2020): "z-Tree unleashed: A novel client-integrating architecture for conducting z-Tree experiments over the Internet," *Journal of Behavioral and Experimental Finance*, 28, 100400.

Dvorak, F., U. Fischbacher, and K. Schmelz (2020): "Incentives for Conformity and Anticonformity," Tech. Rep. TWI Working Paper 122, Available here: https://fdvorak.com/papers/Dvorak-Fischbacher-Schmelz-Incentives-for-Conformity-and-Anticonformity.pdf.

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

Efferson, C., S. Vogt, and E. Fehr (2020): "The promise and the peril of using social influence to reverse harmful traditions," *Nature Human Behaviour*, 4, 55–68.

Fischbacher, U. (2007): "Z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10, 171–178.

Gaertner, W. (1974): "A Dynamic Model of Interdependent Consumer Behavior," *Journal of Economics*, 34, 327–344.

Johnson, N., S. Kotz, and N. Balakrishnan (1994): *Continuous univariate distributions. 2nd ed.*, John Wiley and Sons, chap. Beta distributions, 221–235.

Jones, S. R. G. (1984): *The Economics of Conformism*, Oxford: Blackwell.

Kőszegi, B. (2006): "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, 4, 673–707.

Luce, R. D. (1959): "On the possible psychophysical laws," *Psychological Review*, 66, 81–95.

Manski, C. F. (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 60, 531–542.

McElreath, R. (2020): *rethinking: Statistical Rethinking book package*, r package version 2.13.

McElreath, R., A. V. Bell, C. Efferson, M. Lubell, P. J. Richerson, and T. Waring (2008): "Beyond existence and aiming outside the laboratory: estimating frequency-dependent and pay-off-biased social learning strategies," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363, 3515–3528.

McElreath, R., M. Lubell, P. J. Richerson, T. M. Waring, W. Baum, E. Edsten, C. Efferson, and B. Paciotti (2005): "Applying evolutionary models to the laboratory study of social learning," *Evolution and Human Behavior*, 26, 483–508.

McFadden, D. (1974): *Frontiers in Econometrics*, Academic Press, New York, chap. Conditional logit analysis of qualitative choice behavior, 105–142.

Mullen, B., J. L. Atkins, D. S. Champion, C. Edwards, D. Hardy, J. E. Story, and M. Vanderklok (1985): "The false consensus effect: A meta-analysis of 115 hypothesis tests," *Journal of Experimental Social Psychology*, 21, 262–283.

Neal, R. (2011): *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC, chap. MCMC Using Hamiltonian Dynamics, 116–162.

Nolan, J. M., P. W. Schultz, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius (2008): "Normative Social Influence is Underdetected," *Personality and Social Psychology Bulletin*, 34, 913–923, pMID: 18550863.

Pollak, R. A. (1976): "Interdependent Preferences," *The American Economic Review*, 66, 309–320.

R Core Team (2021): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Revelt, D. and K. Train (1998): "Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level," *The Review of Economics and Statistics*, 80, 647–657.

Sasaki, S. (2019): "Majority size and conformity behavior in charitable giving: Field evidence from a donation-based crowdfunding platform in Japan," *Journal of Economic Psychology*, 70, 36–51.

Schelling, T. C. (1971): "Dynamic models of segregation," *The Journal of Mathematical Sociology*, 1, 143–186.

Schultz, P. W., J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius (2007): "The Constructive, Destructive, and Reconstructive Power of Social Norms," *Psychological Science*, 18, 429–434.

Smerdon, D., T. Offerman, and U. Gneezy (2020): "Everybody's doing it: on the persistence of bad social norms," *Experimental Economics*, 23, 392–420.

Smith, A., B. D. Bernheim, C. F. Camerer, and A. Rangel (2014): "Neural Activity Reveals Preferences without Choices," *American Economic Journal: Microeconomics*, 6, 1–36.

Smith, S., F. Windmeijer, and E. Wright (2015): "Peer Effects in Charitable Giving: Evidence from the (Running) Field," *The Economic Journal*, 125, 1053–1071.

Stan Development Team (2020): "RStan: the R interface to Stan," R package version 2.21.2.

Team, S. D. (2021): *Stan Modeling Language Users Guide and Reference Manual, 2.28*.

Toyokawa, W., A. Whalen, and K. N. Laland (2019): "Social learning strategies regulate the wisdom and madness of interactive crowds," *Nature Human Behaviour*, 3, 183–193.

Vehtari, A., A. Gelman, and J. Gabry (2017): "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and Computing*, 27, 1413–1432.

Watanabe, S. (2010): "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory ." *Journal of Machine Learning Research*, 11, 3571–3594.

Zimmermann, F. (2020): "The Dynamics of Motivated Beliefs," *American Economic Review*, 110, 337–61.

Appendix

Figure 6: Decision screen with slider

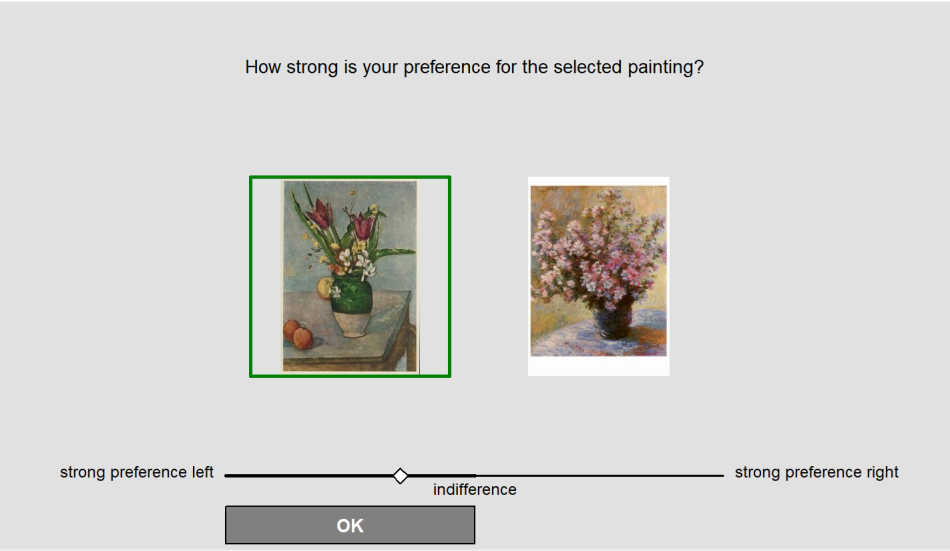


Figure 7: Evaluation screen