From whom do we learn: An exploration of biases in social learning

Yuan Chang Leong (ycleong@stanford.edu)
Department of Psychology, Stanford University
450 Serra Mall, Stanford, CA 94305

Abstract

Decision-making often involves integrating information accumulated from our past experience (*private information)* with information provided to us by other agents (*public information)*. How do we know how much to weight public information? We propose that people weight public information according to its relative accuracy, which can be in turned learned via trial-and-error. However, this learning can be unduly influenced by cognitive biases, leading to overly optimistic estimates of the accuracy of public information. Here, we explore two such biases - optimistic initial expectations and confirmation bias. We formalized these biases in a probabilistic model and demonstrate that model behavior successfully reproduced behavior of human participants in a social learning task.

Keywords: social learning, biases, probabilistic models

Introduction

In studying social influence and herding behavior, economists make the distinction between *private information* that represents one's own beliefs about the state of the world and *public information* that is provided by others (Chamley, 2003). Decisions are then made based on a weighted combination of private and public information. We often consider public information when making decisions because we believe that others have access to relevant information that could be useful to us (Boyd & Richerson, 1985). By combining what we know with what we learn from others, we can make more accurate judgments (Bahrami et al., 2010), choose actions with higher expected values (Biele, Rieskamp, & Gonzalez, 2009; Li, Delgado, & Phelps, 2011) and avoid costly mistakes without having to have made the mistakes ourselves (Olsson & Phelps, 2007). To make adaptive decisions, one should only incorporate public information in so far as it provides useful information about the state of the world. How do people learn which sources of public information are informative? Previous work suggests that human participants are able to track the accuracy of sources of information via trial-and-error, and that they adaptively modulate their use of public information based on the estimated accuracy of the source (Behrens, Hunt, Woolrich, & Rushworth, 2008; Boorman, O'Doherty, Adolphs, & Rangel, 2013).

Yet, the real world is rife with examples of people relying on inaccurate public information – from investors who continue to put their faith in financial advisors despite the advisors being at chance with market predictions (Engelberg, Sasseville, & Williams, 2011), to the average consumer adopting health practices recommended by televised medical talk shows with questionable track records (Korownyk et al., 2014). Even in tightly controlled laboratory experiments, human participants ignored actual environmental contingencies and repeatedly followed misleading advice, leading to suboptimal decisions and outcomes (Biele, Rieskamp, & Gonzalez, 2009; Doll et al., 2009; Staudinger & Büchel, 2013).

If people are indeed able to track the accuracy of informants, why are they so susceptible to inaccurate information from bad informants? In this paper, we build on the models developed in previous work to formally explore the biases that might be present in social learning. We focus on learning from other people, since that is arguably the most common form of social learning (Boyd & Richerson, 1985). We propose two forms of bias – *optimistic initial expectations* and *confirmation bias*. Models of social learning usually assume uniform priors, implying that participants had no information about the sources of information at the beginning of the experiment. Research in social psychology, however, suggests that our initial expectations of people tend to be optimistic (Sears, 1983; Stevens & Fiske, 1995), thus it is possible that participants come to the experiment with optimistic priors of how accurate human advisors would be. Bias can also occur during the updating of beliefs when provided with new information. Confirmation bias is a particularly pervasive bias wherein people interpret new evidence as supporting existing beliefs, even when the evidence is ambiguous or contradictory (Nickerson, 1998). We introduced these biases into a probabilistic learning model, and show that model behavior successfully reproduced human behavior in a social learning task.

Computational Model

Consider an advisor whose accuracy, a, determines whether he makes a correct prediction, y:

$$p(y=1)=a \qquad \text{and} \qquad p(y=0)=1-a \tag{1}$$

We formalized the learning of the advisor's accuracy as a problem of conditional inference – given how accurate the advisor has been in the past, what is the likelihood that he will be accurate in the future. Bayes rule is used to compute the probability distribution of the advisor's accuracy, a, over the course of the experiment given the past successes, y, of the advisor:

$$p(a_{1:i}|y_{1:i}) \propto p(a_1)\prod_{j=1}^{i} p(y_j|a_j) \tag{2}$$

Integrating over the history of *a* up to trial *i*, we obtain the posterior distribution of *a* at trial *i*:

$$p(a_i|y_{1:i}) \propto \int p(a_1) \prod_{j=1}^{i} p(y_j|a_j) da_{1:i-1} \quad (3)$$

Optimistic Prior

We assumed an optimistic prior:

$$p(a_1) = \beta(4,3) \quad (4)$$

such that,

$$E(a_1) = \frac{4}{7} \quad (5)$$

Confirmation Bias

We implemented confirmation bias by adding an additional update rule. On each trial, with probability α, the model interprets new evidence as in line with participants' overall belief about the advisor's accuracy:

With probability α,

if $E(a_i) > 0.5, y_i = 1$

else, $y_i = 0$

Otherwise,

$y_i =$ 1, if advisor made an accurate prediction;

$y_i =$ 0, if advisor made an inaccurate prediction

where $\alpha = |E(a_i) - 0.5|$

Since α scales with the extremeness of the belief, the more accurate or inaccurate the model's estimate of the advisor's accuracy, the more likely it is to invoke the confirmation bias rule.

Action Selection

In our task, participants have to bet for or against the advisor's prediction on each trial. We assume that participants make their bets following a "softmax" policy:

$$p(bet = FOR) = \frac{1}{1 + e^{-\beta(E(a_i)-0.5}} \quad (6)$$

where β is an inverse temperature parameter that models the degree of stochasticity in participants' bets.

Methods

We now describe a behavioral experiment designed to study how human participants learn about private information, how they learn about the accuracy of social sources of information, and how they integrate private information with public information to make decisions. The paradigm is modeled after making investment decisions with advice from financial advisors, and is adapted from tasks used in earlier work on social learning and decision-making (Behrens et al., 2008; Boorman, O'Doherty, Adolphs, & Rangel, 2013).

Participants

26 participants were recruited from the Stanford community. Participants were either Stanford undergraduates who received credit for completing the study or members of the Stanford community who were paid for their participation.

Procedure

The task consists of three phases (Fig. 1). In the first phase of the task (*Private Phase)*, participants were tasked to predict the price fluctuation of the stock. The true probability that the stock price goes up, *pUP*, drifts slowly from trial-to-trial. Participants had to infer *pUP* on each trial based on the past history of the stock. In the second phase of the task (*Social Phase)*, participants evaluated three other social targets (henceforth, advisors) performing the task. On each trial, they had to predict if the advisor will accurately predict the price fluctuation of the stock. Once they made their prediction, they were shown the advisor's prediction, followed by the actual performance of the stock. The advisors were 75%, 50% and 25% accurate respectively. Participants do not know these accuracies beforehand, and had to figure it out via trial-an-error to perform well on the task. In the final phase (*Joint Phase*), participants were again tasked to predict the price fluctuation of the stock. However, prior to making a prediction, they were shown the prediction of one of the advisors whom they encountered in the social phase.

For the purpose of this paper, I will focus on data from the social phase. This is motivated in part because behavior in the private phase has been well studied (Behrens, Woolrich, Walton, & Rushworth, 2007), and in part because a good model for how participants learn the accuracy of the advisors in the social phase will be useful in building a model of their behavior in the joint phase.

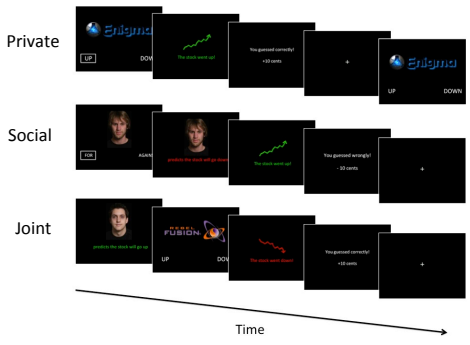

Figure 1: **Schematic of Task**

Model-fitting

The only free parameter in our model is the softmax inverse temperature β. For each participant, we fit the model to choice data to find the optimal value of β that minimized the negative log likelihood of the choices given the model. Model fitting was done using the fmincon function in MATLAB.

Results

Behavioral Results

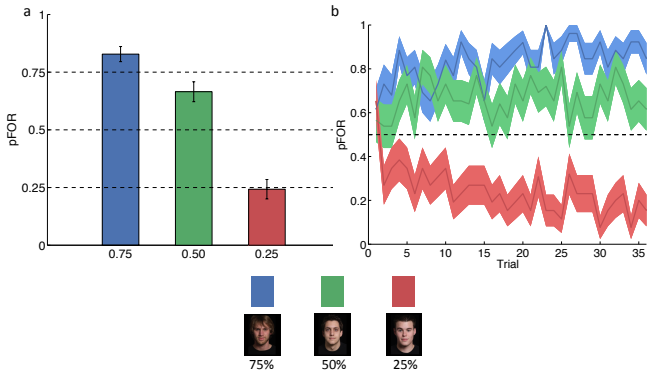

Figure 2: **Behavioral results.** Proportions of trials on which participants bet for each advisor's prediction (a) on average, (b) as a function of trial number.

Participants observed and evaluated three financial advisors predicting the price fluctuations of a particular stock. Here, we were interested in how participants tracked the accuracies of the advisors based on the advisor's past successes and failures at predicting the stock. On each trial, participants had to bet for or against an advisor's prediction. These bets were indicative of how accurate participants thought each advisor was on each trial.

Over the course of the experiment, participants became more likely to bet for the 75% advisor, indicating that they figured out that this advisor was often accurate. Participants also became more likely to bet against the 25% advisor, indicating that they figured out that this advisor was often inaccurate. Participants consistently bet for the 50% advisor, suggesting that they thought the at-chance advisor was in fact better than chance. On average, participants bet for the 50% advisor on more than 50% of the trials ($t(25) = 3.8$, $p < 0.001$, $M = 0.67$, $SE = 0.04$, Fig. 2). When asked to estimate each advisor's accuracy after the experiment, participants also overestimated the accuracy of the 50% advisor ($t(25) = 4.5$, $p < 0.001$, $M = 58\%$, $SE = 2\%$).

Participants were more likely to bet for the 75% advisor than they were to bet against the 25% advisor ($t(25) = 2.3$, $p = 0.03$, $M = 0.16$, $SE = 0.05$). The proportion of bets for the 75% advisor was significantly higher than 75% ($t(25) = 2.4$, $p = 0.02$, $M = 0.83$) but the proportion of bets for the 25% advisor was not significantly different from 25% ($t(25) = -0.18$, $p = 0.86$). The 25% advisor was as inaccurate as the 75% advisor was accurate, and participants ought to bet against the 25% advisor as much as they bet for the 75% advisor. Participants did not do this, suggesting that there might have been asymmetries in how participants tracked the accuracy of accurate versus inaccurate advisors.

Overall, participants' bets revealed that they had a tendency to be overly optimistic about the advisors' accuracy. To understand the biases that could explain the optimistic estimates, we modeled the bets on a trial-by-trial basis.

Modeling Results

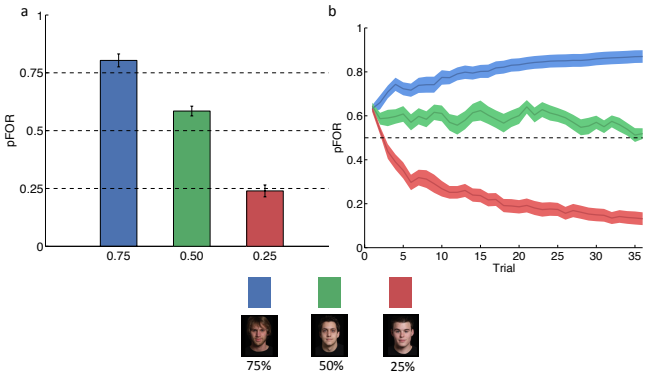

Figure 3: **Modeling results.** Proportions of trials on which the model bet for each advisor's prediction (a) on average, (b) as a function of trial number.

For each participant, we simulated the model performing the same sequence of trials, given the optimal parameter obtained by fitting the data. The model simulations replicated the pattern of results observed in the actual data. The model learned to bet for the 75% advisor and against the 25% advisor. Moreover, the model was consistently more likely to bet for the 50% advisor. On average, the model bet for the 50% advisor more than 50% of the time ($t(25) = 3.99$, $p < 0.001$). The model was also more likely to bet for the 75% advisor than to bet against the 25% advisor ($t(25) = 3.95$, $p < 0.001$).

Discussion

When we make decisions, we often have to combine what we know with what others tell us. To made good decisions, we have to discern between good advice and bad advice. One metric by which we can evaluate the advice we receive is the past accuracy of the information source. In this study, we investigated if and how individuals tracked the accuracy of social informants. We framed accuracy tracking as a problem of probabilistic inference based on observations of past successes and failures. Within this framework, we had participants observe and evaluate the accuracy of financial advisors predicting the price fluctuations of a stock. While participants learned to distinguish between advisors of different accuracies, there were systematic biases in their beliefs about the advisors' accuracies. In particular, participants thought an at-chance advisor was better than chance, and bet for his prediction more than 50% of the time. Furthermore, they were more likely to bet for an accurate advisor than they were to bet against an equally inaccurate advisor

We formally explored the nature of the learning biases using a computational model. We built on previous work that modeled decision-making using variants of the current task (Behrens et al., 2008, 2007; Boorman et al., 2013, Waskom et al., under review). Unlike previous models, our

model assumed optimistic priors, implying that participants started the experiment with optimistic initial expectations about how accurate the advisors would be. We also included a confirmation bias update rule, such that the model probabilistically interprets new evidence as supporting the existing belief, even when the evidence was in fact contradictory. Model simulations successfully replicated the observed pattern of behavioral results, suggesting that optimistic initial expectations and confirmation bias when integrating new information could potentially account for participants' behavior.

Previous work in social psychology suggests that individuals often have moderately optimistic initial expectations of others (Sears, 1983; Stevens & Fiske, 1995), which provides a potential explanation for the optimism in our task. Another possible explanation is that our participants were applying Grice's cooperative principle (Grice, 1975), and assumed that advisors would only give advice if they had access to relevant information. Further work is needed to uncover the origins of optimistic priors. One promising direction for future work is to consider priors on the accuracy of different social targets as reflecting participants' implicit attitudes (Greenwald, Andrew, Uhlmann, & Banaji, 2009; Greenwald, McGhee, & K, 1998). It would be interesting to test if the optimism priors obtained in our task are related to other measures of implicit attitudes, and if manipulating the social category of the advisors would affect participants' priors. In the model, confirmation bias interacts with the optimistic priors such that the optimism persists despite moderate amounts of negative evidence. In other words, the pattern of behavior results in our task was likely due to both biases in priors as well as biases in how new information was integrated. This is consistent with previous work suggesting that confirmation bias can slow down learning and lead to inaccurate inferences about environmental contingencies (Doll et al., 2009, 2009; Staudinger & Büchel, 2013).

Optimistic initial expectations and confirmation bias affect different components of the model – one acts on priors while the other acts on the updating process. By taking a computational modeling approach, we can be quantitatively precise about the nature and magnitude of the biases. This has several advantages, one of which is to provide experimenters a formal framework to study these biases. For example, while the current model imposed a prior on the data, it is possible to infer the biases based on the data. Future work investigating the influence of social categorization on priors can take advantage of this to compare the difference in inferred priors in different conditions. Similarly, we currently assume that each participant has the same degree of confirmation bias. We can add a free parameter to the confirmation bias update rule, which would measure the degree of confirmation bias exhibited by each participant. Furthermore, these models make quantitative predictions about latent variables of participants' cognitive processes. These estimates can then be regressed against neural data to identify corresponding neural mechanisms. Future work can hope to flesh out the cognitive and neural basis of the biases identified in the current study.

Acknowledgements

I would like to thank Noah Goodman for teaching the class that made this project possible, as well as for helpful advice and comments specific to this project. Thanks also to the TAs Desmond Ong and MH Tessler for their help and patience over this past quarter (especially when my assignments were an hour late). I would also like to thank Jamil Zaki for the supervision on the overall project that this paper is part of. I considered adding him to the author list, but it feels weird putting your advisor's name on your homework assignment.

References

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085. http://doi.org/10.1126/science.1185718

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. http://doi.org/10.1038/nature07538

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. http://doi.org/10.1038/nn1954

Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational Models for the Combination of Advice and Individual Learning. *Cognitive Science*, *33*(2), 206–242. http://doi.org/10.1111/j.1551-6709.2009.01010.x

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The Behavioral and Neural Mechanisms Underlying the Tracking of Expertise. *Neuron*, *80*(6), 1558–1571. http://doi.org/10.1016/j.neuron.2013.10.024

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

Chamley, C. (2003). *Rational Herds: Economic Models of Social Learning*. Cambridge University Press. Retrieved from http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/052153092X

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–94. http://doi.org/10.1016/j.brainres.2009.07.007

Greenwald, A. G., Andrew, T., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive

validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. http://doi.org/10.1037/a0015575

Greenwald, A. G., McGhee, D. E., & K, L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. http://doi.org/10.1037/0022-3514.74.6.1464

Li, J., Delgado, M. R., & Phelps, E. A. (2011). How instructed knowledge modulates the neural systems of reward learning. *Proceedings of the National Academy of Sciences*, *108*(1), 55–60. http://doi.org/10.1073/pnas.1014938108

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. http://doi.org/10.1037/1089-2680.2.2.175

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience*, *10*(9), 1095–1102. http://doi.org/10.1038/nn1968

Sears, D. O. (1983). The person-positivity bias. *Journal of Personality and Social Psychology*, *44*(2), 233–250. http://doi.org/10.1037/0022-3514.44.2.233

Staudinger, M. R., & Büchel, C. (2013). How initial confirmatory experience potentiates the detrimental influence of bad advice. *NeuroImage*, *76*, 125–133. http://doi.org/10.1016/j.neuroimage.2013.02.074

Stevens, L. E., & Fiske, S. T. (1995). Motivation and Cognition in Social Life: A Social Survival Perspective. *Social Cognition*, *13*(3), 189–214. http://doi.org/10.1521/soco.1995.13.3.189