The Art of (Recursive Bayesian) Persuasion

Samuel A. Barnett (samuelab@princeton.edu)
Department of Computer Science, Princeton University
Princeton, NJ 08540 USA

Abstract

An honest speaker that is attempting to convince you of some fact must balance the need to be persuasive with the need to appear as an impartial informer. An effective listener will be able to make inferences from a biased speaker if also modeling the fact that the speaker has a bias of some nature. Here, we employ a recursive Bayesian model of two particular phenomena that arise in this context: the weak and strong evidence effects. We show that the model is capable of capturing these effects in an intuitive way, and that the higher-order levels of reasoning are *essential* for doing so.

Keywords: communication; computational modeling; persuasion; theory of mind

Introduction

> There are, then, these three means of effecting persuasion [...] (1) to reason logically, (2) to understand human character and goodness in their various forms, and (3) to understand the emotions
>
> *Aristotle*

> Well he would [say that], wouldn't he?
>
> *Mandy Rice-Davies*

Much of the information we learn comes through communication with actors in the world with their own desires, and epistemological viewpoint. An honest agent that is attempting to convince you of some fact must balance the need to be persuasive with the need to appear as an impartial informer. An effective listener will be able to make inferences from a biased speaker, provided that he is able to take into account the nature of the speaker's biased.

This paper models two particular phenomena that occur in the context of communication with biased agents: the *weak* and *strong evidence effects*.

The *weak evidence effect* typically occurs in the context in which a listener first hears one side of a dispute, followed by the other side (McKenzie, Lee, & Chen, 2002). For example, in a courtroom, jurors first hear the plaintiff's case, followed by the defendant's case. If the plaintiff makes a strong case, then a subsequently weak case made by the defendant may make the jurors *more* confident in the plaintiff's case. This is intuitive from a pragmatic perspective: the jurors expect both the plaintiff and defendant to present the strongest arguments for their case, and so the weakness of the defendant's case suggests an absence of stronger evidence. However, on a naïve model in which we do not consider how the evidence is selected, it would be inconsistent to reduce one's confidence in the defendant's case on hearing her argument, if the argument on its own raises the probability of her side being true (Fernbach, Darlow, & Sloman, 2011).

The *strong evidence effect*, in contrast, shows how stronger evidence may not always lead to stronger inferences (Perfors, Navarro, & Shafto, 2018). In particular, knowing that a speaker's utterances will be used to make inferences about bias can lead to a speaker modulating the evidence she presents to a listener, rather than presenting the strongest possible argument for her case. This has been found to occur even when the distribution of the underlying evidence leans overwhelmingly towards the speaker's side of the argument.

This paper shows how both the strong and weak evidence effects can be predicted and explained under one unified Bayesian computational model in the toy context of *the stick task*. This is the first computational model to demonstrate the strong evidence effect, and the first model of the weak evidence effect employing an explicit theory of mind in its explanation.[1]

The core of this model is recursive Bayesian inference about underlying states given utterances by Bayesian decision-makers, in the style of Rational Speech Act (RSA) models (Goodman & Frank, 2016). In particular, we begin at the lowest level with a naïve judge who assumes the evidence he receives is impartial, and add alternate layers of speakers and judges in order to capture higher-order reasoning about the best means of persuasion. The agenda of each speaker can then be captured by a simple, scalar bias parameter.

This work is also related to work about learning from examples showing that different assumptions about how the data are sampled can have a large impact on learning. In particular, previous models have looked at *weak*, *strong*, and*pedagogical* sampling (Hsu & Griffiths, 2009; Shafto, Goodman, & Griffiths, 2014; Tenenbaum, 1999; Tenenbaum & Griffiths, 2001), each corresponding to different ways in which the data are generated, and subsequently corresponding to different likelihood functions assumed by the learner. This work can be considered to introduce a fourth sampling assumption, *rhetorical sampling*, in which the data are generated by a speaker who is trying to convince the learner of her

[1] Previous work models the weak evidence effect by evaluating case strength with respect to a malleable reference point.

point of view, which also appears in the learner's likelihood function.[2]

The results from this paper show that the recursive Bayesian model of persuasion is able to capture both the weak and strong evidence effects. Moreover, the higher-order models of speakers and listeners are shown to be *essential* for capturing the effect, thus justifying the framework under which this model is designed. Finally, we show that varying parameters such as the speaker bias, or the way in which this bias is perceived by the judge, changes the nature of the weak and strong evidence effects in intuitive ways.

Modeling Evidence Effects

World Model: The Stick Task

The stick task consists of a judge, and a set of speakers indexed by I. A sample of N sticks whose lengths are given by

$$\mathcal{S}_N = \{S_1, S_2, ..., S_N\} \tag{1}$$

are drawn i.i.d. from the Uniform$[0,1]$ distribution and is fixed throughout the task.[3]

Each speaker observes the full sample before the task, and at each time step t a speaker i (one per time step, taking it in turn) chooses one stick from the sample to reveal to the judge. The speakers are not permitted to reveal a stick that has previously been shown to the judge. Notationally, the speaker chooses action

$$a_t^{(i)} \in \{s_1, s_2, ..., s_N\} \setminus \mathcal{A}_{t-1}, \tag{2}$$

where s_i is the realization of random variable S_i and $\mathcal{A}_{t-1}$ denotes the first $t-1$ actions chosen. For simplicity, we let $\mathcal{A}_0 = \emptyset$.

At each time step t, the judge reasons about the sample mean of the sticks, $\bar{S} = \frac{1}{N}\sum_{n=1}^{N} S_n$.[4] In particular, the judge evaluates his posterior over whether the sample is 'long' or 'short', i.e., whether or not $\bar{S}_N \geq 0.5$. Hence, the relevant posterior for the judge at time step t is

$$p(\bar{S}_N \geq 0.5 \mid \mathcal{A}_t). \tag{3}$$

Each speaker will have an incentive to select evidence that persuades the judge that the sample is either long or short, or the speaker will be indifferent towards the outcome.

The task runs for a total of $T \leq N$ time steps.

Agent Models

In the manner of Rational Speech Act (RSA) models (Goodman & Frank, 2016), we model the speakers as performing recursive Bayesian inference about one another's beliefs. To capture both evidence effects, we require four layers, which we divide into the 'naïve' layers and the 'pragmatic' layers. The model is implemented in WebPPL (Goodman & Stuhlmüller, 2014), a probabilistic programming language that allows for fast hierarchical inference in low-dimensional domains such as this one.

Naïve Judge and Speaker The first layer, J_0, describes the naïve judge - this judge is naïve in the sense that he does not model the speakers as having any incentives, and instead assumes that the actions are selected uniformly from the available sample. Relabeling without loss of generality, the posterior of J_0 at time t is given by

$$p_{J_0}(\bar{S}_N \geq 0.5 \mid S_1 = s_1, S_2 = s_2, ..., S_t = s_t). \tag{4}$$

The second layer, S_1, describes the naïve speaker, whose choice about which stick to show at time t is represented as a posterior over the available sticks defined in reference to p_{J0}. Importantly, each speaker has a bias β, where in our simulations we model the biases as being in the range $\beta \in \{-10, -5, -2, 0, 2, 5, 10\}$. This bias represents the incentive regarding the judge's inference: a positive (resp. negative) bias entails that the speaker is incentivized to show sticks to the judge that steer the judge towards the belief that the sample is long (resp. short).

To produce this behavior, the speaker samples a stick based on the soft-maximization of biased informativity of that stick, taking into account the previous sticks shown:

$$p_{S_1}(a_t \mid \mathcal{A}_{t-1},\ \mathcal{S}_N) \propto \exp\big(\beta \cdot a_t \cdot p_{J_0}(\bar{S}_N \geq 0.5 \mid \mathcal{A}_{t-1} \cup \{a_t\}\big). \tag{5}$$

Observe that, if $\beta = 0$, this entails that the speaker will sample the sticks uniformly based on the available evidence, which is equivalent to the speaker being indifferent to the outcome.

Pragmatic Judge and Speaker The pragmatic judge, J_1, performs a similar inference as J_0, though with one crucial difference: the judge models each stick as having been sampled by a naïve speaker. To see how this is possible, observe that we can express the action of speaker i at time t by the random variable

$$A_t^{(i)} \sim p_{S_1}(\cdot \mid \mathcal{A}_{t-1},\ \mathcal{S}_N), \tag{6}$$

which is a categorical distribution computed by Equation 5. For ease of notation, we write the set of observations of S_1 speakers

$$\mathcal{A}'_T = \{A_1^{(i_1)} = a_1^{(i_1)}, A_2^{(i_2)} = a_2^{(i_2)}, ..., A_T^{(i_T)} = a_T^{(i_T)}\}, \tag{7}$$

and let $\mathcal{A}'_0 = \emptyset$.

The posterior of J_1 can now be computed via Bayes' rule:

$$\begin{aligned} p_{J_1}(\mathcal{S}_N \mid \mathcal{A}'_T) &\propto p(\mathcal{A}'_T \mid \mathcal{S}_N) p(\mathcal{S}_N) \\ &= p(\mathcal{S}_N) \prod_{t=1}^{T} p_{S_1}(A_t^{(i_t)} = a_t^{(i_t)} \mid \mathcal{A}'_{t-1}) \end{aligned}$$

[2] A related model in which the agents are allowed to *deceive* a listener is considered in Oey, Schachner, and Vul (2019).

[3] In practice, we use a discrete approximation to this distribution by modeling the sticks as being drawn uniformly from the set $\{0.025, 0.075, ..., 0.975\}$.

[4] It is assumed that the judge knows N, but only observes the stick values through the speakers' actions.

Using the sum rule, we arrive at the relevant posterior, which in the case of discretized stick lengths is given by

$$p_{J_1}(\bar{S}_N \geq 0.5 \mid \mathcal{A}'_T) = \sum_{S_N\,:\,\bar{S}_N \geq 0.5} p_{J_1}(S_N \mid \mathcal{A}'_T). \qquad (8)$$

We consider two distinct versions of the pragmatic judge: one in which he knows the biases of each speaker in advance, and one in which these biases are drawn, i.i.d., from a categorical prior over the aforementioned range of bias values. This is necessary as the weak and strong evidence effects demand the former and latter versions, respectively.

We further divided the latter case into two by considering two possible priors: a flat prior over the bias range, and a 'V-shaped' prior which down-weights the likelihood of neutrality.[5] We considered these two priors as, although the V-shaped prior better describes the context of the stick task, we wished to show that strong evidence effects were robust to our choice of prior. Factoring in uncertainty over the bias is then a matter of marginalizing over the possible bias settings using the sum rule.

The final layer, the pragmatic speaker S_2, is nearly identical to the naïve speaker S_1, except for the fact that the pragmatic speaker performs soft-maximization based on J_1's judgment, as opposed to J_0's judgment:

$$p_{S_2}(a_t \mid \mathcal{A}_{t-1},\, S_N) \propto \exp\big(\beta \cdot a_t \cdot p_{J_1}(\bar{S}_N \geq 0.5 \mid \mathcal{A}_{t-1} \cup \{a_t\}\big). \qquad (9)$$

This layer is only required to capture the strong evidence effects.

Simulations

Simulation 1: Weak Evidence Effect

Set-Up In the first simulation, we captured the pragmatic judge's ability to capture the *weak evidence effect*. We present the judge with one stick in turn from two speakers, whose biases are known to the judge and fixed at

$$\beta_1 \in \{2,5,10\}, \quad \beta_2 = -\beta_1. \qquad (10)$$

We then contrast the naïve and pragmatic judge's posteriors over whether the sample is long, having seen both sticks.

In this context, the weak evidence effect occurs when the naïve judge *decreases* his belief that the sample is long between observing the first and the second stick, while the pragmatic judge *increases* his belief. To see why this is the case, recall that the weak evidence effect is a result of a conditional probability being judged lower than the marginal, while the cause is probability-raising. In the stick task, we evaluate whether the cause is probability-raising with respect to a naïve judge, who effectively regards the sticks as being sampled i.i.d., whereas the 'conditional' and 'marginal' in question (that is, the conditional probabilities after the first and second stick observation) are evaluated with respect to the pragmatic judge with a more nuanced sampling assumption.

We also measure the *strength* of the weak evidence effect for each stick pairing, given by the changes in belief from observing the first and second stick, summed over the naïve and pragmatic judge.[6] In particular, this strength is given by

$$\sum_{J \in \{J_0, J_1\}} |p_J(\bar{S}_N \geq 0.5 \mid \mathcal{A}'_2) - p_J(\bar{S}_N \geq 0.5 \mid \mathcal{A}'_1)|. \qquad (11)$$

For this simulation, we vary the total sample size N between 3 and 5, and we consider the full range of possible stick pairs to be presented.

Results The heatmaps show that our model is capable of capturing the weak evidence effect, with the effect being at its strongest for $a_1^{(1)} \approx 0.675$ and $a_2^{(2)} \approx 0.475$. Notably, this effect seems to *weaken* for stronger values of $a_1^{(1)}$: this is possibly due to the fact that such strong evidence makes it very unlikely that the sample is long, so that the weaker evidence causes a smaller shift in belief for both judges.

We also observe that increasing the magnitude of the biases for both speakers both the range of stick pairs for which we observe the weak evidence effect, as well as the strength of the effect. That this occurs is no surprise: in the limit, observations of sticks from biased speakers effectively informs us of the upper and lower bounds of the sample, giving us a significant amount of information about the sample mean in the case in which the second stick provides 'weak' evidence for the sample being short.

Finally, we see that increasing the sample size broadens the range of stick pairs for which we observe the weak evidence effect.

Simulation 2: Strong Evidence Effect for Speakers

Set-Up In the second simulation, we investigate the strong evidence effect for *speakers*, and in particular, whether adding the pragmatic speaker is truly necessary to capture this effect. We consider a simple speaker with a bias $\beta \in \{2,5,10\}$. The stick task is then run for $T = 2$ steps, with $N = 5$ available sticks. In contrast to Simulation 1, the pragmatic judge (as modelled by the pragmatic speaker) no longer knows the speaker bias, and instead has either a flat or a V-shaped prior over possible bias values.

For each setting of bias value and bias prior shape, we consider 150 draws of initial stick samples (recording the true sample mean) for the speaker to choose from. We then look at the distribution over possible ordered pairs of sticks that the speaker chooses to show the judge, which we refer to as 'strategies'. For computational reasons, this distribution is inferred via a Monte Carlo estimate with 100 samples, as opposed to using the enumeration method that governs the lower-order inferences.

To study the effect, we compare this distribution of strategies to what we call the 'optimal strategy': to show the sticks

[5] Concretely, the V-shaped prior is given by the normalized probability vector $\frac{1}{29}[8,4,2,1,2,4,8]$.

[6] This is set to zero for stick pairings for which the weak evidence effect does not occur.

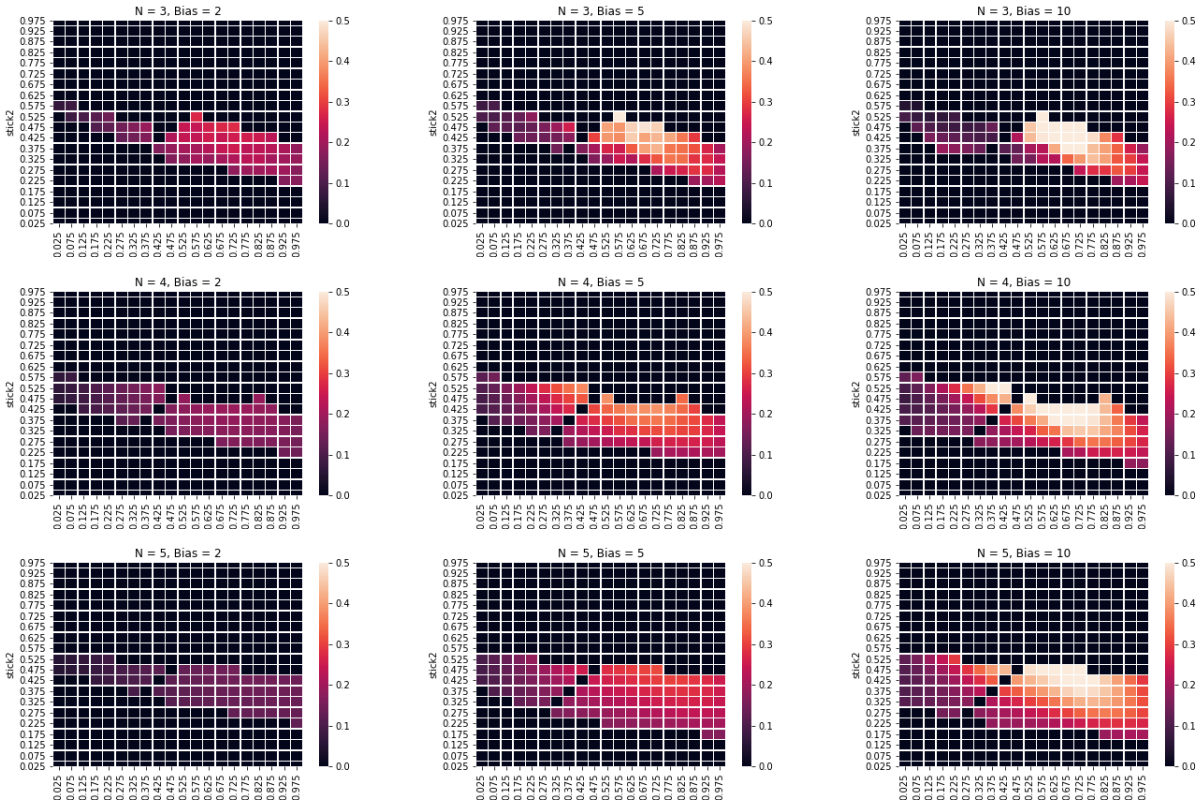

Figure 1: Heatmaps depicting the stick pairs for which the weak evidence effect appears, and the strength of the effect in the cases where it does appear. Values for which there is no weak evidence effect are colored in black. We vary both the sample size N and the agents' bias strength β.

in descending order. Given only a naïve judge, this is the strategy that maximizes the judge's posterior probability that the sample is long. Hence, as $\beta \to \infty$ we would expect the probability mass of this strategy to tend to 1 for the naïve speaker. We therefore compare the probability mass of this strategy for the naïve speaker to the pragmatic speaker, to see if the taking into account that the judge is modeling the speaker bias reduces the likelihood that the speaker will adopt this strategy, and instead leads to a different Maximum A Posteriori (MAP) strategy. Such a reduction would show that the pragmatic speaker demonstrates a strong evidence effect.

Table 1: Percentage of trials for which the MAP strategy is the optimal strategy, given for both speakers S_1 and S_2.

bias prior shape	agent bias (β)	MAP(S_1) = opt (%)	MAP(S_2) = opt (%)
flat	2.0	41.3	37.3
	5.0	76.7	66.7
	10.0	97.3	78.7
V-shaped	2.0	42.7	45.3
	5.0	76.0	69.3
	10.0	97.3	72.0

Results Plotting the probability mass on the optimal probability for the naïve speaker against the pragmatic speaker, we see that, in the majority of cases (**give a proportion**), the naïve speaker has a higher probability of choosing the optimal strategy than the pragmatic speaker. In particular, the probability of choosing the optimal strategy for the pragmatic speaker never exceeds ε higher than the probability for the naïve speaker, and can (and often is) much lower. This shows that the pragmatic speaker is able to capture the strong evidence effect.

Interestingly, we see little difference in whether the judge is modeled to have a flat prior or a V-shaped prior over speaker biases, although the strong evidence effect is slightly more prevalent for flat priors. This may be because, if the speaker believes the judge to already be disposed towards believing that the speaker is biased, then it may be true for the speaker that choosing her sticks non-optimally has less of an impact on the judge's belief about her bias, and so the optimal strategy may look slightly more favorable.

We also see that, for both the naïve and pragmatic speaker, the probability of choosing the optimal strategy *increases* as speaker bias increases and *decreases* as sample mean increases. The former phenomenon is clearly true for the naïve speaker, whereas for the pragmatic speaker we can argue that increasing the bias begins to crowd out considerations about how the judge perceives that bias. The latter phenomenon can

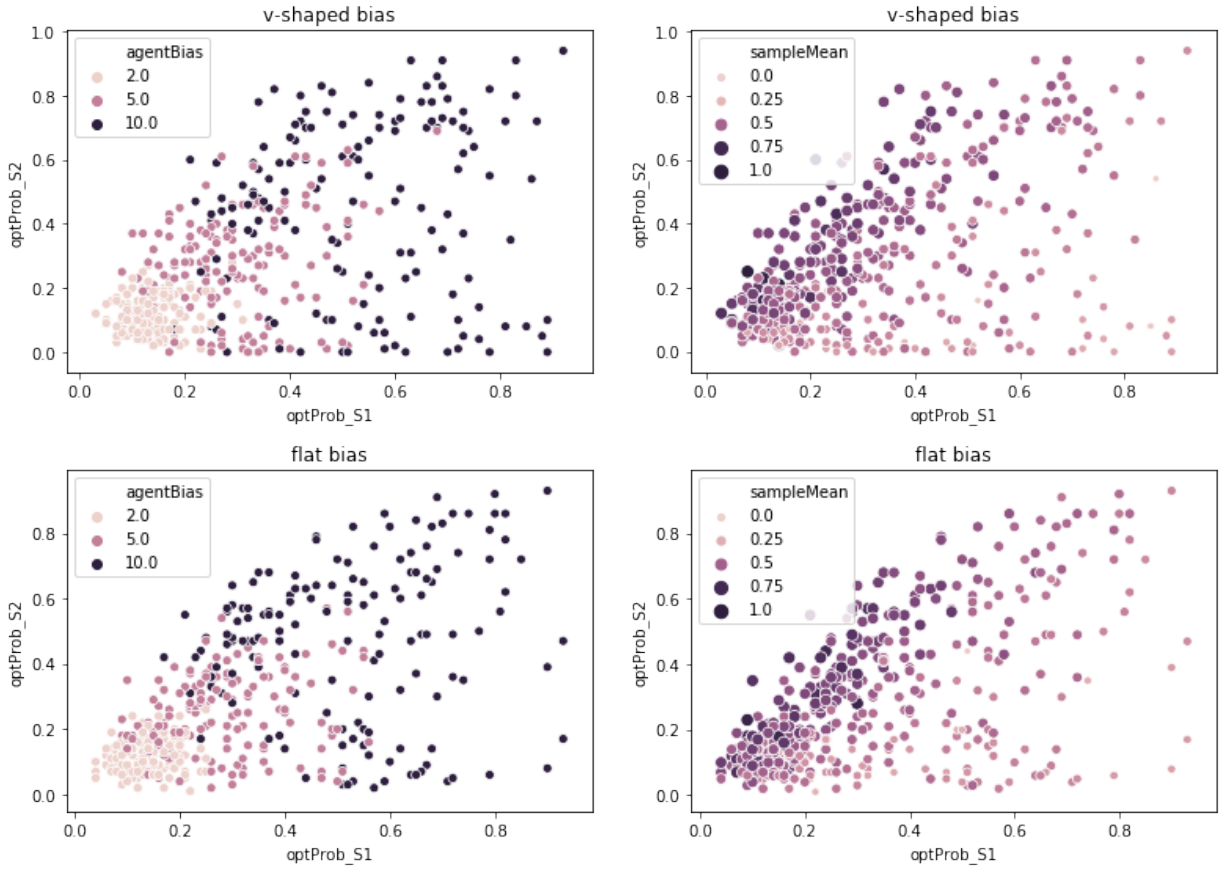

Figure 2: Scatter plots of the probability mass on the optimal strategy for S_1 vs. S_2. Plot points are colored both by sample mean and by agent bias, and results are shown for S_2 modeling a judge with either a flat or V-shaped prior over agent bias.

be explained as follows: in the domain in which the sample mean is high, the speaker can afford to choose a sub-optimal strategy and still get her desired outcome, though if the sample mean is low it becomes increasingly important to just show the longest sticks.

The strong evidence effect also appears by looking at the MAP strategies in Table 1, which shows that (with one exception), the number of inferred distributions for which the optimal strategy was the MAP strategy decreases as we move from a naïve to a pragmatic speaker. Moreover, this gap increases as we increase the bias of the speaker.

Simulation 3: Strong Evidence Effect for Judges

Set-Up In this simulation, we investigate the strong evidence effect for *judges*, which predicts a relationship between the strength of the evidence presented by a speaker, and the judge's perception of the speaker's bias. We again consider one speaker, and look at the judge's posterior probability of the speaker's bias value after observing one stick, where this distribution is computed exactly via enumeration.[7] We also vary whether the judge begins with a flat prior over bias values, or a V-shaped prior.

Results Figure 3 shows, predictably, that an increase stick length leads to an increased posterior probability in higher bias values, whereas for lower bias values this increase eventually plateaus or slightly decreases. This is clearer in the case of a flat bias prior than a V-shaped prior, where the initial skew in belief towards greater bias values suppresses the belief that the speaker is positively biased for lower stick values. However, in both cases the overall shape and trend of the curves are the same.

Discussion and Future Work

Knowing that the information you receive is coming from partisan sources might invite a degree of skepticism, often to the extent that the evidence received provides no informational content whatsoever. Yet by employing a recursive theory of mind, we show that we are able to capture the nuanced influences perceived bias can have, both on speakers and listeners. Our model provides a computational justification for some of the seemingly counter-intuitive behaviors exhibited in rhetorical, argumentative contexts.

Future work would investigate the extent to which RSA-style models can work in higher-dimensional reasoning tasks,

[7] Clearly, the judge does not know *a priori* what the speaker's bias value is in this context, in contrast to Simulation 1.

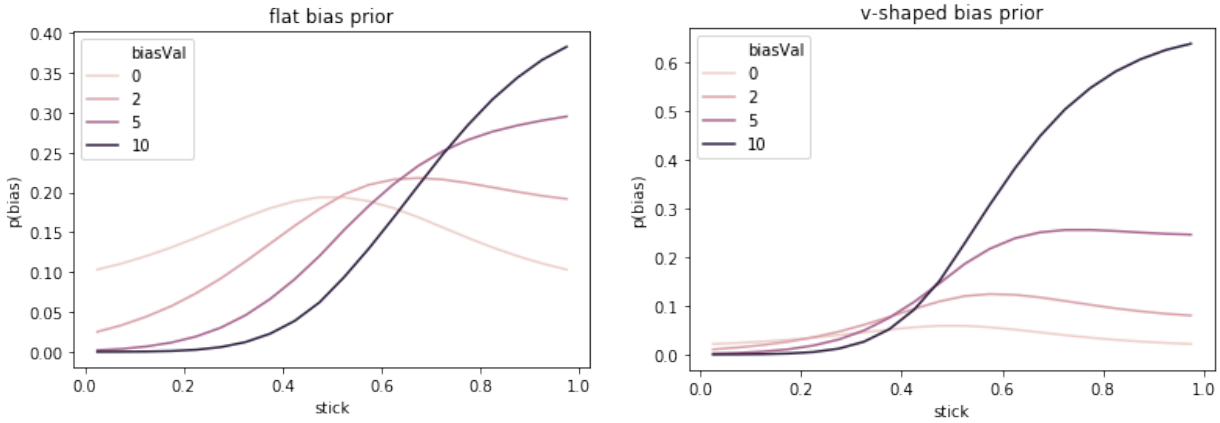

Figure 3: Pragmatic judge J_1's posterior over agent bias, after seeing one stick, where the stick length and initial prior shape are varied.

and whether higher-order speaker models are able to capture other persuasive techniques if given a greater variety of possible utterances (Cialdini, 1993; Falk & Scholz, 2018).

Finally, it would be interesting to apply this work to the recently-proposed context of AI safety via debate (Irving, Christiano, & Amodei, 2018), in which agents compete in a debate game to produce the most true, useful information for a human to judge in a decision-making task. While previous models have looked at agents playing a zero-sum game using Monte Carlo Tree Search, it is plausible that an agent model with a theory of mind about the target of persuasion can perform better in terms of producing useful information. This would be in line with other research indicating the importance of human-like AI models for human-AI interaction (Carroll et al., 2019; Hilgard, Rosenfeld, Banaji, Cao, & Parkes, 2019).

References

Aristotle. (1991). *The art of rhetoric*. New York, N.Y.: Penguin Books.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., & Dragan, A. (2019). On the utility of learning about humans for human-ai coordination. In *Advances in neural information processing systems* (pp. 5175–5186).

Cialdini, R. B. (1993). *Influence: the psychology of persuasion*. New York: Morrow.

Falk, E., & Scholz, C. (2018). Persuasion, influence, and value: Perspectives from communication and social neuroscience. *Annual Review of Psychology*, *69*(1), 329–356.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). When good evidence goes bad: The weak evidence effect in judgment and decision-making. *Cognition*, *119*(3), 459–467.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* http://dippl.org. (Accessed: 2020-1-7)

Hilgard, S., Rosenfeld, N., Banaji, M. R., Cao, J., & Parkes, D. C. (2019). Learning representations by humans, for humans. *arXiv:1905.12686 [cs, stat]*. Retrieved 2019-10-10, from http://arxiv.org/abs/1905.12686

Hsu, A., & Griffiths, T. L. (2009). Differential use of implicit negative evidence in generative and discriminative language learning. In *Advances in neural information processing systems* (pp. 754–762).

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv:1805.00899 [cs, stat]*. Retrieved 2019-05-08, from http://arxiv.org/abs/1805.00899

McKenzie, C. R. M., Lee, S. M., & Chen, K. K. (2002). When negative evidence increases confidence: Change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making*, *15*(1), 17.

Oey, L. A., Schachner, A., & Vul, E. (2019). *Designing good deception: Recursive theory of mind in lying and lie detection* [preprint]. Retrieved 2019-10-25, from https://osf.io/5s4wc doi: 10.31234/osf.io/5s4wc

Perfors, A., Navarro, D. J., & Shafto, P. (2018). Stronger evidence isn't always better: A role for social inference in evidence selection and interpretation. *CogSci*, 6.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In *Advances in neural information processing systems* (pp. 59–68).

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(4), 629–640.