The Art of (Recursive Bayesian) Persuasion

Anonymous CogSci submission

Abstract

An honest speaker that is attempting to convince you of some fact must balance the need to be show the strongest evidence for that fact 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; rational analysis; 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, model of the world, and epistemological access to it. An honest speaker that is attempting to convince you of some fact must balance the need to show you the evidence that best supports her case with the need to appear as an impartial informer. An effective listener will be able to make inferences from a biased speaker, regarding both the bias of the speaker and the content of his utterances.

Tom: Spend a moment on the big picture modeling challenge before getting into specific phenomena.

This paper models two particular phenomena that occur in the context of communication with biased speakers: 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*. Simulation 1 is based on experiments in McKenzie et al. (2002), and Simulations 2 and 3 are based on Studies 1 and 2 in Perfors et al. (2018), respectively. 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.

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 cap-

[1] Previous work models the weak evidence effect by evaluating case strength with respect to a malleable reference point. Furthermore, faint praise, a phenomenon related to the weak evidence effect, has been given a Bayesian formalization by Harris, Corner, and Hahn (2013).

turing 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.

Background

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

Modeling Evidence Effects

World Model: The Stick Task

We model the weak and strong evidence effects on a simple toy task called *the stick task*. In this task, a judge must reason about the mean of a sample originally drawn from a known distribution, where observations from the sample are presented by biased agents.

In particular, 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. Observe that, for the above posterior, the length of each stick observed is directly proportional to the strength of the belief that the sample is long. This direct correlation allows us to clearly model the weak and strong evidence effect by treating the length of each stick as a proxy for the strength of the evidence supporting the statement.

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

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

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. If $\beta \geq 0$, then this is written as:

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

For $\beta < 0$, we simply replace the event $\bar{S}_N \geq 0.5$ in the above expression with $\bar{S}_N < 0.5$. Observe that, if $\beta = 0$, this entails that the speaker will sample the sticks uniformly based

[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.

on the available evidence, which is equivalent to the speaker being indifferent to the outcome.[5]

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

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

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

In order for the pragmatic judge to reason about the naïve speaker, the posterior of J_1 must reflect the agent biases (denoted $\boldsymbol{\beta}$) as well as the distribution over remaining sticks. This can be computed by Bayes' rule:

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

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

$$p_{J_1}(\bar{S}_N \geq 0.5 \mid \mathcal{A}'_T) = \sum_{\boldsymbol{\beta}} \sum_{\substack{\mathcal{S}_N \\ \bar{S}_N \geq 0.5}} p_{J_1}(\mathcal{S}_N \mid \mathcal{A}'_T). \tag{8}$$

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. Given the above notation, we can view the former case as modeling $p(\boldsymbol{\beta})$ as a delta distribution on the known bias values of each speaker.

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.[6] 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 varied with respect to our choice of prior.

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. For $\beta \geq 0$, this is given as

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

with the expression for $\beta < 0$ changing as for S_1.

This layer is only applicable to the simulations capturing 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

$$\beta_1 \in \{2, 5, 10\}, \quad \beta_2 = -\beta_1. \tag{10}$$

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.[7] In particular, this strength is given by

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

For this simulation, we vary the total sample size N between 3 and 5, and we present to the judge the simplex of possible stick pairs where the length of the second stick does not exceed the length of the first.

Results The heatmaps (Fig. 1) 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 $a_2^{(2)} \approx 0.47$. 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 short, 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 increases both the range of stick pairs for

[5] The β parameter also plays the same role as the optimality parameter α in models such as that of Goodman and Frank (2016).

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

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

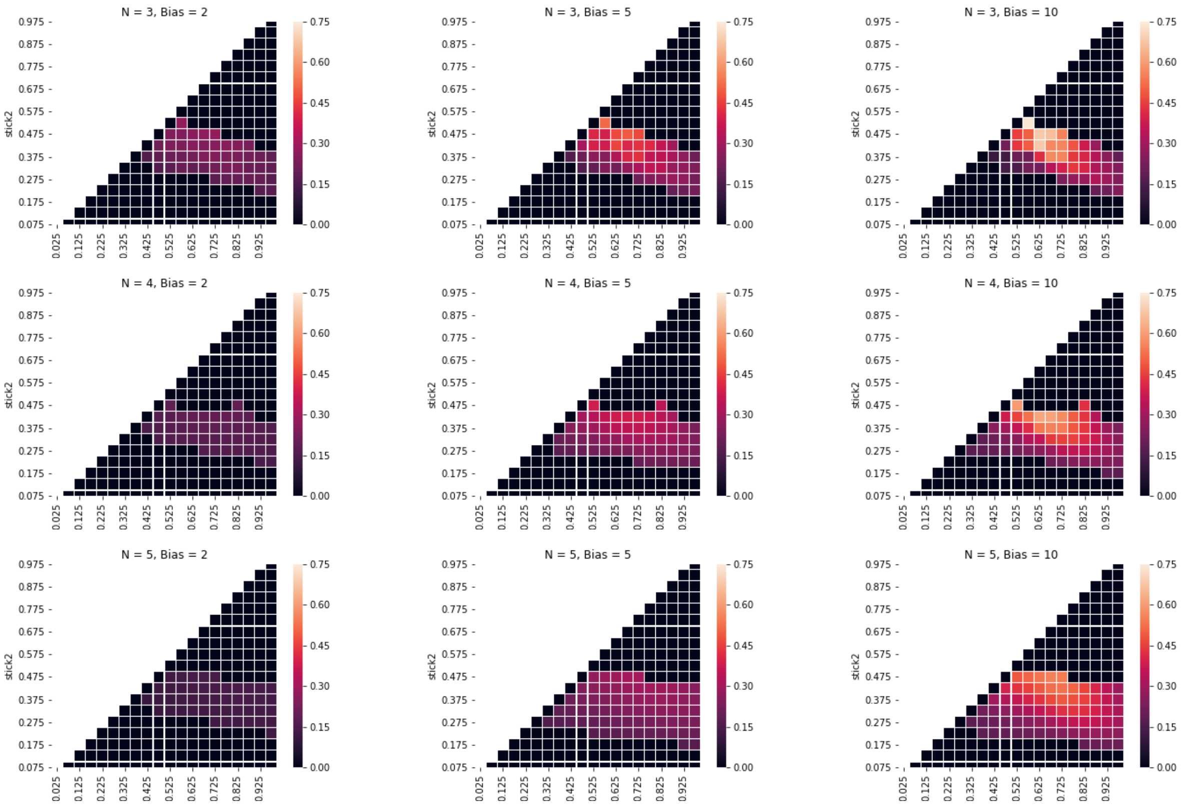

Figure 1: Heatmaps depicting the stick pairs for which the weak evidence effect appears, and the strength of the effect in the cases where it does appear. Values for which there is no weak evidence effect are colored in black. We vary both the sample size N and the agents' bias strength β.

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 inform 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 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.[8] We also vary whether the judge begins with a flat prior over bias values, or a V-shaped prior.

Results Fig. 2 shows, predictably, that an increased 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.

Experiment

To evaluate our model, we implemented a version of the stick task that could be played by a human judge responding to stimuli from two 'contestant' speakers. This allowed us to test whether the weak evidence effect and the strong evidence effect for listeners was present in the stick task, and whether the model could predict the strength of the effect and the circumstances under which it occurs.

Procedure

X participants were recruited via MTurk to participant in our experiment using `nosub`. Participants were briefed that they were to be the judge of a debate between two contestants, one of which would be rewarded \$10 based on the participant's final judgment about the sample. In order to be included in

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

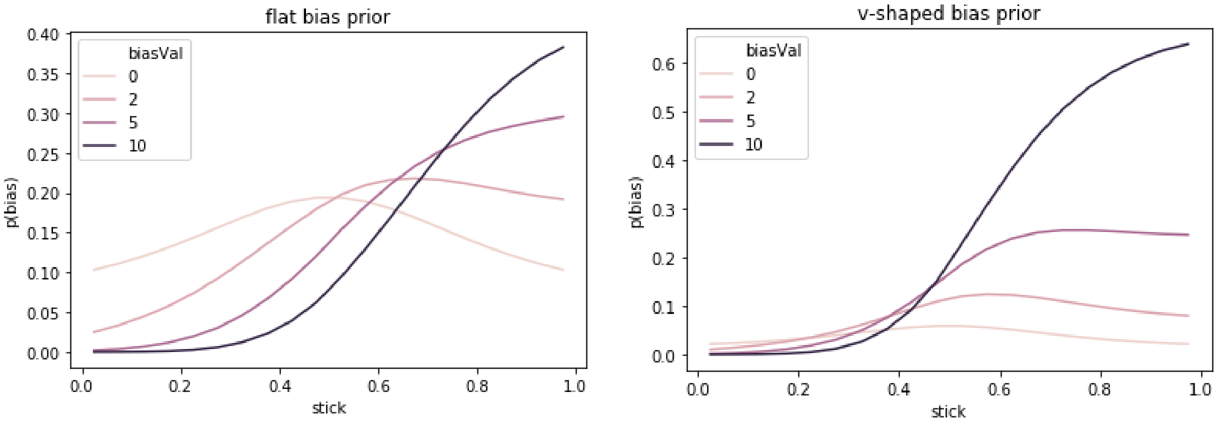

Figure 2: Pragmatic judge J_1's posterior over agent bias, after seeing one stick, where the stick length and initial prior shape are varied.

the trial, they must successfully complete a quiz about the briefing within three attempts.

For one turn each, a 'contestant' would reveal a stick to the participant from an overall sample of 5 sticks: the first (resp. second) contestant shows a stick to the participant with a length in inches in the range $\{6, 7, ..., 10\}$ (resp. $\{1, 2, ..., 5\}$). Each pairing of sticks acts as a cell into which the participants were divided. After being presented each stick, the participant must rate from 0 to 100 their belief that the average of the overall stick sample is longer than 5 inches. The participant is then given a 2AFC task to make a final decision on whether the sample is 'long' or 'short'.

Finally, the participant is asked to rate between 0 and 100 their belief in the bias of each contestant, where 0 (resp. 100) represents the a strong belief that the contestant is biased towards showing shorter (resp. longer) sticks, and 50 means that the contestant is showing sticks at random. The participant is reminded of the length of sticks that each contestant presented, as well as the payment incentive of each contestant.

Results

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 speaker is completely disregarded: hence the disarming power of the expression, "well he would say that, wouldn't he?" Yet by employing a recursive theory of mind, we show that we are able to capture some of the more nuanced influences perceived bias can have, both on speakers and listeners. Our model provides a computational justification for the behaviors exhibited in rhetorical, argumentative contexts that seem counter-intuitive when the listener is treated as a naïve Bayesian with no theory of mind.

Future work would investigate the extent to which RSA-style models can work in higher-dimensional reasoning tasks with a greater number of speakers and listeners, 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). The ultimate aim of such research would be to provide the computational underpinnings of a *rhetorical principle*, by means of analogy to Grice's cooperative principle (Grice, 1975), which would describe how people interact in social settings in which they aim not only to inform but also to *persuade.*.

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 tasks in which humans and AI systems must cooperate (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)

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, speech acts* (Vol. 3). Academic Press.

Harris, A. J. L., Corner, A., & Hahn, U. (2013). James is polite and punctual (and useless): A bayesian formalisation of faint praise. *Thinking & Reasoning*, *19*(3), 414–429.

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.