Understanding *belief bias* by measuring prior beliefs for a Bayesian model of syllogistic reasoning

Michael Henry Tessler
mtessler@stanford.edu

Department of Psychology, Stanford University

Abstract. The phenomenon of *belief bias* in syllogistic reasoning occurs when the a priori believability of a conclusion influences the intuitive acceptability of that conclusion. Prior beliefs about the world can be formalized into a probabilistic generative model of situations. Tessler and Goodman (2014) proposed that this very idea can account for the range of acceptabilities of conclusions from categorical syllogisms with abstract content. Here, I generalize their model to accommodate syllogistic reasoning data where content effects are observed. I collect data about the prior plausibility of various properties co-occurring, and use this data to predict syllogistic reasoning behavior in a separate experiment. I compare models with different types of assumptions concerning the prior and discuss open questions for this approach.

Your logic-chopping friend is in a room with a number of lamps and lightbulbs; you are in a different room and cannot see what she sees. She gives you the following logic puzzle:

> All of the lightbulbs that are hot are bright.
> Some of the lightbulbs that are bright are *not* on.
>
> Are some of the hot lightbulbs *not* on? Are any of the hot ones on?

Prior beliefs about the world guide our actions, thoughts, and reasoning in new situations. It can be helpful, for example, to know how fast a particular kind of animal can run, if you are also thinking about if that animal can eat you. Similarly, humans can use prior beliefs in a domain (e.g. life expectancies) to reason accurately about everyday contexts (e.g. guessing how long someone will live; Griffiths & Tenenbaum, 2006). Finally, it has been argued that prior beliefs influence the very meaning of words (Goodman & Lassiter, 2015). It is odd then that so little formal theory has gone into understanding prior beliefs in classic reasoning tasks (but, cf. Klauer, Musch, & Naumer, 2000; Dube, Rotello, & Heit, 2010).

Bayesian approaches to cognitive science have a natural way of accounting for prior beliefs in reasoning. Tessler and Goodman (2014) described a generative model of argument strength that uses a truth-functional semantics applied to idealized situations composed of objects with properties. This model accounted for much of the variability in Chater and Oaksford (1999)'s meta-analysis data of categorical syllogistic reasoning. That work further explored syllogistic reasoning by incorporating Gricean principles,

formalized in the Rational Speech-Act (RSA) theory of language understanding (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). This pragmatic component was important in capturing important qualitative phenomena in syllogistic reasoning (e.g. the relative preference for the *all X are Y* conclusion over the *some X are Y* conclusion when both are logically valid). This work was done with respect to meta-analysis data that differed largely in the materials used, and for which prior beliefs about the materials were not expected to have a substantial effect. However, it is known that prior expectations about the categories and properties at stake in a syllogism influence the acceptability of a conclusion (J. S. Evans, Handley, & Pollard, 1983; Cherubini, Garnham, Oakhill, & Morley, 1998; J. S. Evans, Handley, & Harper, 2001).

Here, I generalize Tessler and Goodman (2014)'s model of argument strength to capture qualitative phenomena associated with *belief bias*. This is done by empirically measuring prior beliefs about real-world content, deriving model predictions based on those beliefs, and testing the probabilisitic model of argument strength against behavioral data obtained in a separate experiment of syllogistic reasoning. A secondary, primarily methodological concern is about the granularity of information needed to capture these syllogistic reasoning phenomena.

To foreshadow the results, empirically measured priors (Expt. 1) coupled with a Bayesian model of argument strength accounts for much of the syllogistic reasoning data (Expt. 2), including qualitative effects of content. The predictions of the model, which has no parameters, are as good as those of a model with a prior parametrized by 12 variables. The most likely values (conditioned on the data of Expt. 2) of these 12 variables correspond roughly with the marginal distributions of the priors elicited in Expt. 1. This interesting correspondence suggests the syllogistic reasoning task is too coarse-grained to disambiguate models of reasoning that rely on correlated properties from models where independence is assumed.

1 Bayesian argument strength in syllogistic reasoning

A formal account of gradience in syllogistic reasoning was presented by Tessler and Goodman (2014). The computational model is a Bayesian model; as such, it is important to understand the implications of the prior for syllogistic reasoning. I review the model, highlighting along the way how I generalize the model to consider content effects.

1.1 Ontology

The model is based on an ontology of situations composed of objects with properties, similar to mental models (Johnson-Laird, 1983). A situation $s \in S$ is composed of n objects: $s = \{o_1, o_2, ..., o_n\}$, each of which can have 3 properties:

$$s = \{\{A_{o_1}, B_{o_1}, C_{o_1}\}, \{A_{o_2}, B_{o_2}, C_{o_2}\}, ..., \{A_{o_n}, B_{o_n}, C_{o_n}\}\}$$

Properties A, B, and C of these objects are stochastic and assumed to be Boolean for simplicity. Properties *across* objects are assumed to be independent and identically distributed *(iid)*; hence,

$$P(s) = \prod_{1 \leq i \leq n} P(A_{o_i}, B_{o_i}, C_{o_i}) = (P(a,b,c))^n$$

To account for syllogistic reasoning in Chater and Oaksford (1999)'s meta-analysis of 5 studies, which differed with respect to the materials used, the model assumed no *a priori* information about the meaning of the properties; thus, properties *within* objects were determined independently and identically (*i.i.d.*): $P(A_{o_i}, B_{o_i}, C_{o_i}) = P(A_{o_i}) \cdot P(B_{o_i}) \cdot P(C_{o_i}) = (P(p))^3$, with $p \sim \text{Bernoulli}(\theta)$.

The number of objects in a situation n is a parameter of the model, as is the base rate θ of properties. In fitting the model to the meta-analysis data, Tessler and Goodman (2014) found $\theta \approx 0.25$, qualitatively consistent with the "rarity assumption"—that properties are relatively rare of objects—first used by Oaksford and Chater (1994). The best fitting n was around 5, also consistent with the "minimal model assumption" of the Mental Models framework (Johnson-Laird, 1983).

1.2 A generative model of argument strength

The generative model of situations can be turned into a generative model of syllogistic reasoning by providing a semantics for the quantifier sentences of a syllogism. The model uses the interpretation of quantifier sentences as truth-functional operators, consistent with standard practice in formal semantics.

A quantifier utterance (e.g $u_{\text{all } A \text{ are } B}$) maps two properties (e.g. A and B) to a truth value by consulting the properties of the objects in the situation s and applying the usual literal meaning. For instance:

$$\begin{aligned}
[\![u_{\text{no } A \text{ are } B}]\!] &= \{s \in S : \|o_A \cap o_B\| = 0\} \\
[\![u_{\text{some } A \text{ are } B}]\!] &= \{s \in S : \|o_A \cap o_B\| > 0\} \\
[\![u_{\text{all } A \text{ are } B}]\!] &= \{s \in S : \|o_A \cap o_B\| = n\} \\
[\![u_{\text{not all } A \text{ are } B}]\!] &= \{s \in S : \|o_A \cap o_B\| < n\}
\end{aligned}$$

where $o_A = \{o_i | A_{o_i} = 1\}$ and $o_B = \{o_i | B_{o_i} = 1\}$ represent the objects in a situation that have the properties A and B, respectively. Thus, the quantifier utterances pick out the situations s where the truth-functional meaning of the utterance is satisfied.

Truth-functional meanings of quantifier expressions are useful here because an expression which assigns a Boolean value to a situation can be used for probabilisitic conditioning. That is, these quantifier expressions can be used to update a prior belief distribution over situations into a posterior belief distribution:

$$P(s \mid u_1, u_2) \propto P(s) \cdot \delta_{[\![u_1]\!](s)} \cdot \delta_{[\![u_2]\!](s)}$$

where u_1, u_2 are the two quantifier-utterances corresponding to the premises of a syllogism (e.g. $u_{\text{all } A \text{ are } B}, u_{\text{some } B \text{ are not } C}$).

For syllogistic reasoning, we are interested not in the posterior distribution over situations *per se*, but the distribution on true conclusions that these situations entail: $P(u_3 \mid s)$, where u_3 is a quantifier-utterance corresponding to the conclusion of a syllogism (e.g. $u_{\text{some } A \text{ are } C}$). Hence,

$$P(u_3 \mid u_1, u_2) \propto P(u_3 \mid s) \cdot P(s \mid u_1, u_2)$$

This model, thus, returns a posterior distribution over conclusions conditioned on the premises of a syllogism being true.

The Bayesian model has a natural way of accounting for the influence of prior beliefs in reasoning. Indeed, beliefs simply specify a prior distribution over situations. In particular, the assumption that properties in a situation are independent and identically distributed *(i.i.d.)* must be relaxed if we are to consider real-world content. I generalize the model by considering that properties can have correlations; the representation of expectations about the presence or absence of properties will be generalized from one marginal distribution—$P(p) = P(a) = P(b) = P(c)$—to the joint distribution: $P(a,b,c)$.

The model was written in the probabilistic programming language WebPPL[1](Goodman & Stuhlmüller, 2014). For background and details on this form of model representation, see `http://probmods.org`.

2 Experiment 1: Measuring $P(a,b,c)$ for real-world content

Bayesian models of reasoning and language typically measure the relevant prior distribution for a given task in order to generate predictions. For the model of syllogistic reasoning presented by Tessler and Goodman (2014), the natural prior to measure is the distribution over the presence and absence of the properties mentioned in the syllogism. I constructed content domains to intuitively cover a range of probabilities.

Design I recruited 70 participants on Amazon's Mechanical Turk to rate the likelihood of various combinations of properties co-occurring. Participants were paid $0.80 for their work.

To assess the reliability of the elicitation task, I ran the experiment using two different dependent measures as a between-subjects variable. Each participant was randomly assigned to either the "frequency" or the "plausibility" dependent measure condition (described below). Within each of these conditions, participants completed the judgment task for 4 content domains[2].

Procedure & Materials I selected property domains based on model simulations using qualitatively different priors (elicited from people in my lab). These preliminary simulations suggested that domains with causal structure led to the biggest differences between content domains (possibly due to the probability estimates being more reliable for causal domains). Table 1 shows the properties used.

The prompts for the "plausibility" condition read: *Imagine an X (e.g. a lightbulb). How likely is it that it is ___?* The prompts for the "frequency" condition read: *Imagine 100 Xs (e.g. lightbulbs). About how many of them are ___?* Below these prompts were listed the 8 possible combinations of the presence and absence of 3 properties (e.g. *is*

[1] A fully-specified version of this model can be accessed at: `http://forestdb.org/models/syllogisms-esslli2015.html`

[2] The experiment in full can be accessed at: `http://stanford.edu/~mtessler/experiments/syllogism-belief-priors/prior-exp.html`

on, *is bright*, and *is hot*). Next to each set of properties, was a slider bar. In the plausibility condition, the slider bar ranged from "Impossible" to "Certain", with intermediate arrows pointing to the left and right indicating "less likely" and "more likely". In the frequency condition, the slider bar ranged from "0" to "100", with intermediate arrows pointing to the left and right indicated "fewer" and "more".

Experiment 1 Domains				
Noun	Causal relation	Property A	Property B	Property C
crackers	common effect	are soggy	are past expiration date	have lots of flavor
knives	common effect	are sharp	are rusty	cut well
lightbulbs	common cause	are on	are bright	are hot
strawberries	common cause	are in the freezer	are soft	are warm

Table 1. Content domains used in experiments.

Data analysis and results Participants' responses were normalized within each domain so that the ratings for the 8 property combinations made a well-formed probability distribution (i.e. they added up to 1). I then took the mean rating for each of the 8 property combinations in each of the 4 domains, to arrive at mean empirical priors for all 4 domains. These were used as the empirical $P(a,b,c)$ for the Bayesian model.

The experiment elicited unique priors for each domain (see Figure 1). The data elicited with different dependent measures were highly correlated ($r_{pearson} = 0.78$; $r_{spearman} = 0.85$). Though the correlation between the prior data elicited by different dependent measures is good, the data set as a whole was substantially more reliable (95% bootstrapped CI for $r_{split-half} = [0.95, 0.98]$), suggesting meaningful differences between the two measurements. At the same time, model predictions based on the different dependent measures were also substantially more reliable ($r_{pearson} = 0.95$). This suggests that the prior elicitation task captured the relevant variance for the syllogistic reasoning model. For simplicity, I later present the predictions of the reasoning model based on collapsing the prior elicitation ratings across dependent measures, though predictions based on either dependent measure alone are not meaningfully different.

3 Experiment 2: Syllogistic reasoning about real world content

In this experiment, I tested if the real world content from Experiment 1 influenced the conclusions drawn from categorical syllogisms.

Design I recruited 254 participants from Amazon's Mechanical Turk. All participants were required to have a 95% approval rating for their previous work on the web service. Participants were paid $0.60 for their work. Each syllogism was paired with each domain used in Experiment 1. A total of 8 syllogisms were used, resulting in 32 unique $\{$syllogism, domain$\}$ pairs.

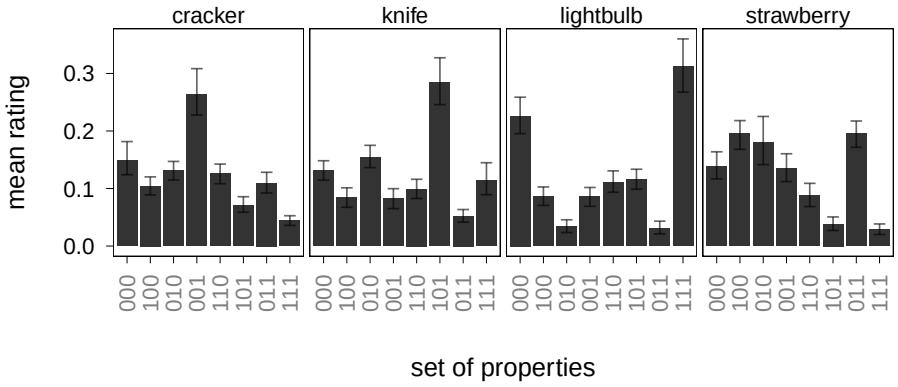

Fig. 1. Mean elicited priors collapsed across dependent measure (see text for details). Error bars denote 95% confidence intervals. X-axis shows presence or absence of each property, the ordering of which can be found in Table 1. For example, the tallest bar in the cracker domain (001) is a cracker which isn't soggy, isn't past expiration date, and has lots of flavors.

Procedures & Materials Each participant completed 4 syllogisms. On each experimental trial, participants were presented with the syllogism (e.g. *Some of the lightbulbs that are bright are on. None of the lightbulbs that are hot are bright.*) and each possible conclusion (e.g. $\{$*All, Some, Not all, None*$\}$ *of the lightbulbs that are hot are on.*) and asked *Does it follow that: X*, for each conclusion. Radio buttons with the options "Doesn't follow" and "Follows" were provided. Below that was a vertically-oriented slider bar with endpoints labeled "Certain" and "Don't know" to measure confidence. Participants were required to mark each conclusion before continuing to the next trial[3].

Results Shown in Figure 2 (lighter bars) are a subset of the results of the experiment. Content effects can be observed by comparing panels within a row (i.e. comparing across columns). For example, for the *some / none* syllogism (top row), the proportion of responses endorsing "Some of the lightbulbs that are hot are on" is appreciably higher than the proportion endorsing "Some of the crackers that have lots of flavor are soggy" (2nd row, columns 1 & 3; *some* conclusion). Effects of reasoning (i.e. of syllogism) can be observed by comparing panels within a column (i.e comparing down rows). For example, "Some of the crackers that have lots of flavor are soggy" is a substantially more endorsed conclusion if the premises are: "All of the crackers that are past expiration date are soggy. Some of the crackers that have lots of flavor are past expiration date." (4th column, rows 1 & 2; *some* conclusion).

[3] The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/syllogism-belief/syllbelief-exp2.html`

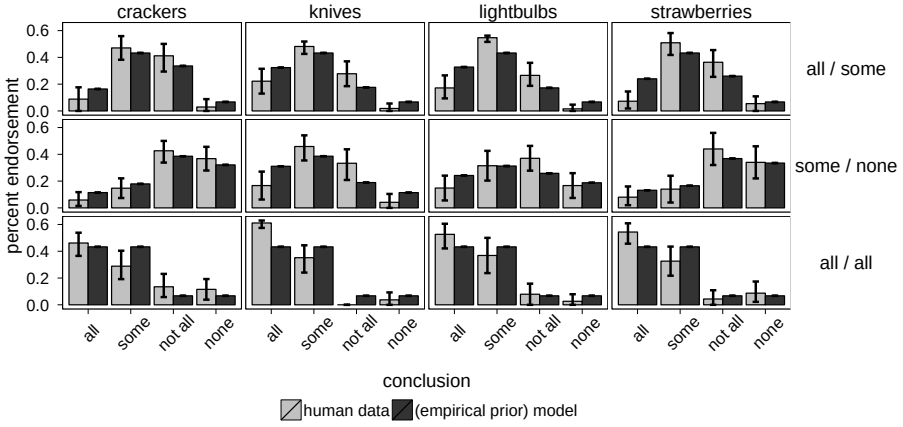

Fig. 2. Reasoning patterns and predictions for 3 (of the 8) syllogisms in the experiment. Human reasoning data is in the darkest shade. The lighter shade is the 12-parameter ("independent") model. The medium shade is the 0-parameter ("empirical prior") model. All syllogisms shown here were of the form B-A / C-B and the conclusions were of the form C-A (e.g. First row: All B are A, Some C are B). Reasoning effects can be seen by comparing rows within a column. Content effects can be seen by comparing columns within a row.

4 Bayesian analysis of Bayesian reasoning models

In this section, I explore 4 different models of argument strength that vary in their independence assumptions in the prior[4]. The first model—the "abstract" model—uses a single base-rate parameter in its prior[5]; this model assumes properties are *i.i.d.* both within domains and across domains. The second model—the "within-*i.i.d.*" model—is the simplest parametrized model that can predict content effects; it is identical to the first model except in that the base-rate parameter can vary across domains (but not within a domain; hence it is *i.i.d.* within but not across domains). This model has 4 parameters (one base-rate for each domain). The third model—the "fully independent" model—assumes only that properties of an object are independent (not necessarily identically distributed); it uses a different base-rate parameter for each property within and across domains. This model has 12 parameters (3 properties per domain and 4 domains in total). The final model is a model that uses the empirically elicited priors from Expt. 1; this model has no free variables parametrizing the prior over properties.

One parameter is shared by all models. This is the number of objects in a situation n, which controls the size of the worlds reasoned over. A preliminary analysis revealed that results were highly consistent for $n \geq 4$ and so I use $n = 4$ for all simulations.

[4] For all of the models I consider, properties are assumed to be independent *across* objects (e.g. o_1 having property A does not influence o_2's chance of having property A). It is the assumption of independence *within* objects that is explored in this paper.

[5] This is the model of argument strength used by Tessler and Goodman (2014) to model meta-analysis data.

I analyze the models by putting uninformative priors ($\theta \sim \text{Uniform}(0,1)$) on the base-rate parameter(s) and conditioning on the observed experimental data. In addition, for all models I include a single data analytic "guessing" parameter $\phi \sim \text{Uniform}(0,1)$ to account for response noise. This noise parameter is important to accommodate data points that deviate largely from what the reasoning model predicts[6].

Inference for the combined Bayesian data analysis of the Bayesian reasoning model was done via the Metropolis-Hastings algorithm implemented in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014). Models were analyzed independently, and MCMC chains were run for 10,000 samples.

4.1 Posteriors over model parameters

The posterior value for ϕ, the guessing variable, was near 0.25 for all models; this analysis attributes about a quarter of the responses to noise. Note that this estimate of noise is with respect to the reasoning model. In other words, it is an estimate of the proportion of responses better explained by random guessing than by the reasoning model. If the posterior predictive distribution of the model predicts the data well, ϕ would be an accurate measure of the response noise in the syllogistic reasoning task. Alternatively, ϕ could also include aspects of the data set that these particular reasoning models do not predict well.

One hypothesis about the base rate parameters of the parametrized-prior models is that the parameter values that account best for the reasoning data are lower-order representations of the empirically-elicited prior. To test this, I compared the posterior distributions over the base rate parameters of the 12-parameter, "fully independent" model to the marginal distributions of the empirical prior data (Figure 3). For the properties corresponding the conclusion terms of the syllogism (properties A & C), the inferred base-rates based on the 12-parameter model are qualitatively consistent with the marginal distributions from Experiment 1. For example, the most likely base rates for property A for the cracker and strawberry domains are definitively smaller than those of the knife and lightbulb domains (Figure 3, column 1, rows 1 & 4 vs. rows 2 & 3). As well, the most likely base rate for property C of the strawberry domain is smaller than the other 3 domains. The inferred base rates from the 12-parameter model have the most uncertainty about property B, possibly because B is only indirectly related to the responses, which are statements about properties A & C. Overall, this is suggestive evidence that the parametrized model of argument strength is using base rates corresponding to those of the marginal distributions over properties elicited in Expt. 1.

[6] In some cases, this parameter is actually *necessary* to analyze the data. This is the case with logically impossible conclusions (e.g. *All A are B // All B are C ∴ No A are C*); in this case, the reasoning model gives this conclusion probability 0 (i.e. *logically impossible* means probability 0). If, for whatever reason, the experimental data includes this conclusion as a response, the data analysis model will crash because that particular data point is expected to have probability 0. The guessing parameter lets us accommodate any data point. This is done by postulating that with ϕ probability, the participant selects a conclusion at random. I put a distribution over the probability and infer this value from the experimental data.

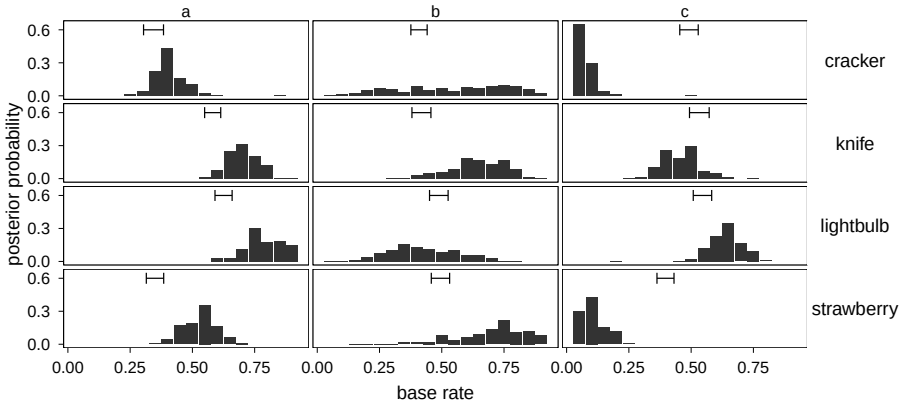

Fig. 3. Marginal posterior distributions over the base rates of properties in the 12-parameter, fully independent model. 95% CIs for the marginal distributions derived from the empirical priors elicited in Experiment 1 are shown at the top of each plot. Qualitative consistencies between the two suggest the base rates inferred from the reasoning data resemble lower-order statistics (the marginal distributions) of the richer joint-distributions shown in Figure 1.

4.2 Posterior predictives

Model	free parameters for priors	other free parameters	correlation with data
abstract	1	1	0.81
within-*iid*	4	1	0.86
fully independent	12	1	0.90
empirical	0	1	0.87

Table 2. Modeling results.

The posterior predictive distribution marginalizes over the inferred parameter values to produce predictions about what the data should look like given the posited reasoning model and the observed data. This is akin to fitting the parameters and is an important step in model validation as it shows what data is actually predicted by the model. All of the models did considerably well in accounting for the variance in the syllogistic reasoning data. Table 2 shows the model–data correlations for each of the models. Figure 4 shows the fit for the empirical prior model.

Figure 2 shows predictions for the empirical prior model (darker shade) together with the experimental data (lighter shade) for 3 example syllogisms. The model based on the empirical prior shows effects of content. For example, the knives and lightbulbs domains show lower endorsements for the *not all* conclusion relative to the crackers and strawberries domains (and the reverse can be observed for the *all* conclusion), consistent with participants' reasoning behavior (Figure 2, row 1: columns 2 & 3 vs. 1 & 4, *not*

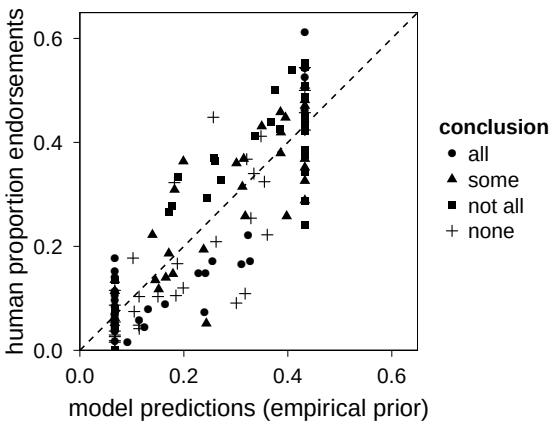

Fig. 4. Data vs. (empirical prior) Model plot. The model provides a good overall fit ($r = 0.87$) to the 128 data points (32 syllogism, domain pairs X 4 conclusions each). The posterior predictions bottom out around 0.07. This is the work of the "guessing" parameter ϕ.

all conclusion). In addition, the model shows effects of reasoning. For example, the endorsement for the *some* conclusion is much higher for the *all / some* premises than for the *some / none* premises, also consistent with the experimental data (Figure 2, column 3: rows 1 vs. 2, *some* conclusion). Further, an interaction between content and reasoning can be observed by comparing the lightbulb domain to the knife domain for those same syllogisms (columns 3 vs. 2 X rows 1 vs. 2). In the knife domain, the experimental data shows the effects of the syllogism are much weaker (the light bars are not very different from one another). The model predictions are also very similar for these two syllogisms in the knife domain.

Finally, it's worth drawing attention to the bottom row of Figure 2, the *all / all* syllogism. This is a valid syllogism (a maximally strong argument) with two logically valid conclusions: *all* and *some*. Participants show a consistent preference for the *all* conclusion over the *some* conclusion, consistent with many other studies of syllogistic reasoning (Khemlani & Johnson-Laird, 2012). It's interesting that this asymmetry is robust across the different content domains. The model predicts consistent responses across the content domains because the syllogistic argument is so strong. However, it has no way to capture the asymmetry between *all* and *some* because it is using only a truth-functional semantics (*all* entails *some*). Tessler and Goodman (2014) predicted this asymmetry by extending to the model to take into account pragmatic reasoning. I leave for future work the incorporation of pragmatic reasoning with syllogistic arguments about real-world content.

5 Discussion

I have demonstrated that a model of syllogistic reasoning with a rich prior distribution can account for the flexibility of interpretation of a syllogistic argument with real world

content. The phenomenon of "belief bias" can be viewed in this framework as a natural extension of the notion of "argument strength". Arguments vary in strength depending on the prior distribution of the properties in question.

This modeling work reveals that the empirical prior model predicts the data well. Using Bayesian data analytic techniques, I observed that the 4-parameter "within-*iid*" model and the 12-parameter "fully independent" model can also accommodate the content effects well. I say the models *accomodate* the data because their predictions are dependent on the particular parameter settings inferred *from that data*. The empirical prior model, by contrast, predicts the data with no parameter fitting. Additionally, it's likely that performing a formal, Bayesian model comparison between these models would favor the empirical prior model due to Bayes' Occam's Razor. However, it is interesting to consider the implications of these modeling results as they stand.

The 4-parameter "within-*iid*" model is the simplest model that could possibly account for content effects. What this model posits is that there is some difference in the base rates of properties in these four different domains. In terms of cognition, this might mean that our artificial domains (e.g. knives that could be sharp, rusty, and/or that cut well) call to mind a general intuition about the base rate of these properties, and that this general base rate enters into the computation of argument strength. The posterior over base rate parameters for this model is consistent with this explanation: base rates for domains with properties that tended to co-occur (e.g. the lightbulbs domain) were relatively high while those for domains with properties that tended *not* to co-occur (e.g. the crackers domain) were low. The 12-parameter "fully independent" model also accommodates the data well. This model posits that properties in a given domain are independent but differ in their base rates. Correlations between properties need not be tracked explicitly.

An alternative explanation for the ubiquitous good fits is that the experiment itself is confounded. The 8 syllogisms used in my experiment might not be the best syllogisms to distinguish models with subtly different independence assumptions about the priors over properties. I found that the most likely base rate parameter values for the parametrized prior models given the data from Expt. 2 were those that roughly corresponded to the marginal distributions of properties from Expt. 1. Both models (parametrized priors and empirical priors) modeled the data equally well, suggesting that these experiments were not well suited to disambiguate them. An even more radical proposal is that categorical syllogistic reasoning *in general* is not the best kind of experiment to distinguish these models. In categorical syllogisms, there are only 3 logical possibilities for conclusions entailed by situations: *all*, *some and not all*, or *none*. It's possible that these models would make different predictions if we allowed more possible responses, e.g. *most* and *few*, exact numbers.

Finally, it is worth noting that classical reasoning experiments such as the one explored here use language to communicate information for participants to reason over. Basic communicative principles, however, are often ignored in the scientist's analysis of behavior. Tessler and Goodman (2014) formalized basic communicative principles in their probabilistic pragmatics model of syllogistic reasoning, as an extension of the Rational Speech-Act theory of language understanding. I leave for future work the interplay between pragmatics and prior beliefs in the syllogistic domain.

References

Chater, N., & Oaksford, M. (1999). The Probability Heuristics Model of Syllogistic Reasoning. *Cognitive psychology*, *258*, 191–258.

Cherubini, P., Garnham, A., Oakhill, J., & Morley, E. (1998). Can any ostrich fly?: some new data on belief bias in syllogistic reasoning. *Cognition*, *69*(2), 179–218.

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: it's a response bias effect. *Psychological review*, *117*(3), 831–63.

Evans, J. S., Handley, S. J., & Harper, C. N. (2001). Necessity, possibility and belief: a study of syllogistic reasoning. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, *54*(3), 935–58.

Evans, J. S., Handley, S. J., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*, 295–306.

Evans, J. S. B. T., Handley, S. J., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*, 295–306.

Frank, M. C., & Goodman, N. D. (2012). Quantifying pragmatic inference in language games. *Science*, *336*, 1–9.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory, 2nd edition.* Wiley-Blackwell.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–84.

Goodman, N. D., & Stuhlmüller, A. (2014). *The design and implementation of probabilistic programming languages.* https://dippl.org.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), 767–73.

Johnson-Laird, P. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological bulletin*, *138*(3), 427–57.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, *107*(4), 852–884.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

Tessler, M. H., & Goodman, N. D. (2014). Some arguments are probably valid: Syllogistic reasoning as communication. In *Proceedings of the 36th annual conference of the cogntiive science society* (pp. 1574–1579).