Contents

List of Figures

The form of logic is conversation

Michael Henry Tessler, Noah D. Goodman

manuscript draft

> Reasoning draws the conclusion and makes us grant the conclusion, but does not make the conclusion certain, nor does it remove doubt so that the mind may rest on the intuition of truth, unless the mind discovers it by the path of experience; since many have the arguments relating to what can be known, but because they lack experience they neglect the arguments, and neither avoid what is harmful nor follow what is good...
>
> **Roger Bacon**, *Opus Majus* (1267)

> The syllogism is like a text: fixed, boxed-off, isolated... The riddle [by contrast] belongs in the oral world. To solve a riddle, canniness is needed: one draws on knowledge, often deeply subconscious, beyond the words themselves in the riddle.
>
> **Walter J. Ong**, *Orality and Literacy* (1982)

The syllogism is regarded as the first formal system of logic. Syllogism means roughly *form of logic*

Imagine you're discussing the 2014 World Cup with your friend. Your friend remarks:

> "*All* of the teams that made it to the semifinals were expected to advance. But also, did you notice that *some* of the teams expected to advance had favorable referees?

You may conclude that probably some of the semifinalists had favorable referees previously (and probably because they were expected to advance). The argument seems reasonable, and yet it isn't valid in the deductive sense. Why does it seem reasonable?

1 Introduction

Your friends argument sounds like a syllogism: a formal logical argument that uses quantifiers (e.g. *all*, *some*) to relate two properties (e.g. being a *semifinalist* and having *favorable referees*) via a middle property (e.g. being *expected to advance*).

All A are B	**A – B**
Some B are C	**B – C**
—	—
Some A are C	**A – C**

Table 1: World cup syllogism, without content [left] and without quantifiers [right]

For ease of representation, the syllogism can be represented without the content (Table 1, left) and without the relations between terms (i.e. just as a logical form; Table 1, right).

The syllogism is considered the first formal system of reasoning, developed by Aristotle in the 4th century BC. Syllogistic reasoning has been replaced by more modern formalisms (e.g. the predicate calculus), yet persists in the experiments of cognitive psychologists. Syllogisms are an intriguing testing ground for human reasoning because they (a) are a formal system of logic and (b) use natural language in their construction. The former ensures there is a normatively correct solution (i.e. the syllogism is either logically valid or it is not), while the fact that they use natural language allows people without explicit training in formal logic to analyze these arguments.

It is a fact that most syllogistic arguments are invalid: there is no relation between A & C which can be deduced with absolute certainty (meaning, there is no relation between A & C that *is true in every situation in which the premises are true*). When people are presented with these arguments, however, they are very often comfortable drawing some conclusion. A recent meta-analysis of syllogistic reasoning performance showed that over the population, the appropriate production of a *no valid conclusion* response to an invalid syllogism ranged from 76% to 12%. For valid arguments, the accuracy of producing appropriate, valid conclusions ranged from 90% to 1% (Khemlani & Johnson-Laird, 2012): drawing deductively valid conclusions doesn't come naturally to people.

1.1 Sketch of paper

In this paper, we consider human reasoning performance on the syllogistic reasoning task as Bayesian inference. We formalize a probabilistic model of syllogistic reasoning that reasons over abstract *situations*, consisting of objects with properties. These abstract situations are similar in sprit to mental models applied to syllogisms (P. Johnson-Laird, 1983). The model is generative and thus makes forward predictions about the distribution of responses for each syllogism. In addition, our approach is fundamentally Bayesian and the prior distribution has a serious interpretation. It has been known for some time that the content of syllogisms affect reasoning (?, ?) but never before has a model been able to accommodate background knowledge without positing additional processing steps and/or

some assumed categorical distinction between believable and unbelievable statements. Background knowledge fits naturally within the Bayesian model framework, as it simply describes an empirical prior distribution over situations. We test the influence of the prior distribution over situations (i.e. background knowledge) on conclusions drawn in 4 syllogisms where our model predicts differences. The model predicts subtle graded effects of background knowledge and the experiment confirms these predictions.

We further scale the Bayesian syllogism model using a generic prior to account for a number of data sets in the literature (Chater & Oaksford Meta-analysis, Khemlani & Johnson-Laird meta-analysis, Rips, ...). The overall correspondence to these data is strong, suggesting the gradedness in human reasoning can be considered as probabilistic inference. The Bayesian model fails, however, to account for a robust qualitative phenomenon in syllogistic reasoning: a strong preference for a single conclusion in syllogisms with two logically valid conclusions.

This phenomenon cannot be understood by the semantics of language alone. We go beyond natural language semantics, and use recent developments in the field of probabilistic pragmatics to consider pragmatic effects in the syllogistic reasoning task. We extend the Rational Speech-Acts (RSA) framework (?, ?, ?) to the syllogistic reasoning task. Standard RSA is insufficient, however. The model replicates the finding that standalone scalar implicature (i.e. that the quantifier "some" may in fact imply "not all") does not explain syllogistic reasoning tasks.

If we consider syllogistic reasoning as a sort of conversation, however, then we must decide what is at stake in the conversation. In other words, what is the *Question Under Discussion*, or the QUD. We replace the standard RSA QUD ("what is the world like?") with a syllogistic QUD ("what is the relationship between the conclusion terms"?). This leads to different pragmatic inferences than standalone scalar implicature. Importantly, this breaks the symmetry between equally valid conclusions. We show how the pragmatics QUD model provides an overall better fit to the large published data sets.

We go on to ask how background knowledge interacts with pragmatic inference. We found that the pragmatics model makes the intriguingly strong prediction of an asymmetry reversal for certain syllogisms with certain background knowledge. Experiment 2 results did not confirm this prediction. We consider the possibility that reasoners rethink what is in common ground when faced with particularly odd premises. We formalize this in an alternative pragmatics model that reasons over both the conclusion implied by the argument and *what is assumed to be in common ground*.

This model predicts no reversal of the asymmetry under very strong background knowledge (because in essence, when the premises are thought to be extremely unlikely a priori, the reasoner thinks the experimenter is thinking of a different situation than what the typical definition would imply). We tested the predictions of these two alternative models in a second experimenter and found that the model that does simultaneous inference over which world and which conclusion both

captures the qualitative phenomena as well as provides a better quantitative fit to the data.

Finally, our model is based on a truth-functional semantics and as such, it is able to accommodate any quantified sentence with a truth-functional meaning. We extend the model by looking at 2 syllogistic reasoning data sets that use the quantifiers *most* and *few*.

This is not the first model to report high correlations between the model predictions and the observed data. We analyze the adequacy of our models with respect to each other as well as to a dominant computational model from the literature: the Probability Heuristics Model. We do this by drawing upon the tools of Bayesian Data Analysis. Bayesian model comparison takes into account not only the fit of a model to the data set but also what other data sets the model could have fit, utilizing Bayes' Occam's Razor, the notion that "a model that explains everything, explains nothing". We explore the implications of the Bayesian Data Analysis.

2 The variety of syllogistic reasoning tasks

Syllogistic reasoning has been a topic of immense interest in cognitive psychology for over a hundred years (Störring, 1908). Since then, there have been dozens of studies of syllogistic reasoning, each with a slightly different design and goal in their specification.

The tasks have taken on a number of forms in their long history. We will first review the space of reasoning problems and then discuss each of the three main tasks in the literature.

2.1 The syllogistic space

Testing reasoning using syllogisms is an attractive approach because there is a fully enumerated space of problems, what we'll call *the syllogistic space*. Indeed, it is this feature of syllogisms that has led some scientists to consider syllogistic reasoning as the premier test case for cognitive science as a whole (P. Johnson-Laird, 1983). Until that point, psychologists had focused on the errors in reasoning rather than the mental representations and processes. Khemlani and Johnson-Laird (2012) reviews 12 different theories of syllogistic reasoning, concluding that none provides a satisfactory account of the phenomena.

The syllogistic space is defined by taking all possible combinations of premise term orderings (Table 2) and quantifiers {*all*, *none*, *some*, *not all*}. The space consists of 64 pairs of premises. For each premise-pair, there are 8 possible conclusions: 4 quantifiers in both the A–C / C–A order. This yields a space of 512 syllogisms (P. N. Johnson-Laird & Steedman, 1978).

Note that two of the four quantifiers {*some*, *none*} represent symmetric relations: e.g. *some A are C* is semantically equivalent to *some C are A*, though there may be interesting pragmatic differences between these (?, ?). It may be of interest as well that only the first three of the term orderings (called "the figures") were considered by Aristotle. The fourth was added by his pupil

B – A	**A – B**	**B – A**	**A – B**
C – B	**C – B**	**B – C**	**B – C**
—	—	—	—
A – C	**A – C**	**A – C**	**A – C**

Table 2: The 4 unique term-orderings ("figures") of syllogisms

Theophrastus, and was contended by some of the scholastic logicians (e.g. Peter Abelard) to not constitute a unique argument form (Lagerlund, 2012). Indeed, if conclusions may proceed in either the A–C or C–A construction, it is a completely redundant form.

If we remove the redundant forms (due to both the symmetry of *some* and *none* and the fact that Figures 1 & 4 are logically equivalent when A–C and C–A conclusions are allowed) , we are left with a space of 32 premise-pairs, each with 6 unique conclusions. This is a substantially smaller space of arguments (192) than what has been considered in the past to constitute the full syllogistic space (512 arguments) (P. N. Johnson-Laird & Steedman, 1978). That is not to say that there are not meaningful differences between the 320 syllogisms with their logically redundant counterparts. It is to say, however, that these differences should not be attributed to differences in logic of the argument.

2.2 Forced choice task

The earliest usage of syllogisms in experimental psychology research used a *forced choice* paradigm. Subjects would be presented with the premises of the argument and typically between 3-5 choices of conclusions.

One advantage of using a forced choice paradigm is that it requires subjects to make a decision, even if they are uncertain. A disadvantage of this paradigm is that for several syllogisms more than one valid conclusion exists.

2.3 Production task

In the production task, subjects are presented with the premises of the syllogism and asked: *What follows?* This forces the subject to come up with a conclusion on her own and can lead to many responses that do not fit the form of syllogistic conclusions.

2.4 Evaluation task

In the evaluation task, the subject is presented with the full syllogism: premises and conclusion. The task is usually the to make a forced choice between *valid* and *invalid*. In addition, confidence ratings are used to elicit the subject's uncertainty associated with each response.

3 The argument strength of a syllogism

Our hypothesis is that reasoning with syllogisms can be understood as a special case of language understanding, which itself relies on reasoning in everyday contexts. Everyday reasoning is uncertain and best described by the tools of probability theory. In the probabilistic framework, a deductive argument is understood as an argument that is maximally strong, with a spectrum of argument strength existing below it, given by Pr(*conclusion* | *premises*) (Oaksford & Chater, 2007; Lassiter & Goodman, 2014).

3.1 Probabilistic model

We begin with reasoning that operates over concrete situations. For our purposes, we can idealize these situations as collections of objects with properties (as we'll see, in the syllogism, only three properties — the terms of the syllogism — need be represented). To capture the uncertainty in everyday cognition, we'll be interested in describing distributions over objects with properties. Probabilistic programming languages are a natural formalism to do this. We use the language Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008), a kind of higher-order probabilistic logic based on the lambda calculus(?, ?). For background and details on this form of model representation, see `http://probmods.org`.

Situations are composed of N objects:

```
(define objects (list 'obj1 'obj2 ... 'objN))
```

(Ellipses indicate omissions for brevity, otherwise models are specified via runnable Church code[1].) Syllogisms deal with 3 terms or classes of objects, and so the objects in these situations need only represent 3 properties. Properties `A`, `B`, and `C` of these objects are represented as functions from objects to the property value. For simplicity, we assume properties are Boolean, and so property values can be `true` or `false`. Initially, we assume no *a priori* information about the meaning of the properties; thus, they are determined independently:

```
(define A (mem (lambda (x) (flip br))))
(define B (mem (lambda (x) (flip br))))
(define C (mem (lambda (x) (flip br))))
```

[1] A fully-specified version of this model can be accessed at: `insertlinkhere`

`flip` is what is known as an *Exchangeable Random Primitive*. It is a function that returns a sample from a distribution, in this case, the bernoulli distribution. `br` is the argument to `flip` which corresponds to the Bernoulli parameter p, or success probability, and rages from 0 to 1. Thus, `(flip br)` returns the outcome of a coin flip, and the coin is weighted by `br`; if `(= br 0.5)`, then this amounts to a fair coin flip. Note that the operator `mem` memoizes these functions, so that a given object has the same value each time it is examined within a given situation, even though it is initially a random variable (via `flip`). Previous probabilistic models (Oaksford & Chater, 1994) have invoked a principle of rarity from the observation that properties are relatively rare of objects in the world[2]. For us, this simply means the base rate, `br`, of properties is small.

We interpret the quantifier sentences of syllogistic reasoning as truth-functional operators, consistent with standard practice in formal semantics. A quantifier (e.g. `all`) is then a function of two properties (e.g. `A` and `B`) which maps to a truth value by consulting the properties of the objects in the current situation. For instance:

```
(define all
  (lambda (A B)
    (all-true (map (lambda (x) (if (A x) (B x) true))
                   objects))))
```

Here the helper function `all-true` simply checks that all elements of a list are true, i.e. that all the *As* are indeed *Bs*. The function `map` applies the given function —`(lambda ...)`— to each element of the list `objects`. Similarly we can define `some`, `none`, `not-all` to have their standard logical meanings. We take *existential import* — that sets are non-empty — as an assumption, i.e. *all As are Bs* cannot be true if there are no As. do we need this?

The model then samples a conclusion uniformly from the list of conclusions true of the current situation.

```
(define sample-conclusions (lambda (A C) (uniform-draw (true-conclusions A C))))
```

The key observation to connect truth-functional meanings of quantifier expressions to probability distributions over situations is that an expression which assigns a Boolean value to each situation can be used for probabilistic conditioning. That is, the quantifier sentences can be used to update a prior belief distribution over situations into a posterior belief distribution. For syllogistic reasoning we are interested not in the posterior distribution over situations *per se*, but the distribution on true conclusions that these situations imply. In Church this looks like:

Listing 1: Full probabilistic model

```
(query
 (define objects (list 'o1 'o2 ... 'on))
 . . . define A, B, C . . .
```

[2]This article is an article and it's about reasoning, but it's not a cat, and it's not a car, nor an elephant nor the color red. In fact, there's a very large number of things which this article is not.

```
. . . define all, some, not-all, none . . .
(define conclusion (sample-conclusion A C))

conclusion

(and (premise-one A B)
     (premise-two B C)))
```

`query` is a special function in Church used for probabilistic conditioning. The first arguments to a query function are a generative model, expressed as a series of definitions: background knowledge with which a reasoning agent is endowed. It is here where the uncertainty is expressed: both `conclusion` and `A, B, C` are random choices, as we've seen above. The second argument, called the *query expression*, is the aspect of the computation about which we are interested; it is what we want to know. The final argument, called the *conditioner*, is the information with which we update our beliefs; it is what we know. As the number of situations sampled grows large, the distribution over conclusions converges to the $P(\text{conclusion} \mid \text{premises})$.

4 Correspondence to existing data sets

To test the predictions of the model we used data from a meta-analysis of syllogistic reasoning tasks presented by ? (?) as well as a independent data set presented by ? (?). The meta-analysis data were compiled from five studies on syllogistic reasoning that used both an Evaluation Task or a Forced Choice task (see Section 2 for more details on the differences between these tasks). The data are in the form of percentage response for conclusions that contain each of the 4 quantifiers as well as for "No Valid Conclusion" (NVC). The Bayesian reasoning models described so far are not equipped to handle NVC[3]. We removed the NVC responses from the meta-analysis and renormalized so the endorsement of all conclusions for a syllogism adds to 100. Some studies in the meta-analysis asked participants to draw conclusions which were restricted to the classical ordering of terms (C-A) while others allowed conclusions in either direction (A-C or C-A). To accommodate this, we allowed our model to draw conclusions in either order and collapsed responses across these two orderings to compare it to this data set.

? (?) data are in the form of proportions of endorsements for each conclusion for each syllogism, collected using an Evaluation Task (see Section **??**, for more details on this type of task). To compare to this data set, we modified the cognitive model. Instead of sampling a conclusion from the set of true conclusions, the model evaluated each conclusion and reported whether or not the conclusion followed.

[3]In each possible situation, at least one of the four conclusions will be true. In fact, since the four possible quantifiers form two pairs of logical contradictions, exactly two conclusions will be true in each situation. For example, *all* and *not all* cannot both be true, but one must be true. The same is the case for *none* and *some*.

The models each have two parameters: the number of objects in situations and the base rate of properties. We fit these parameters to the data by maximizing the correlation between the model and the data. We chose to maximize correlation as opposed to likelihood because of the presence of necessary and impossible conclusions. Under our model, logically valid conclusions (and their contradictions) have posterior probability equal to 1 (or 0). Under a likelihood model for fit, we would be discounting certain syllogisms for this very reason.

4.1 Results

For each model, we report the total number of syllogisms for which the model's modal response is the same as for in the meta-analysis data. This is a qualitative assessment of fit. The table below shows the number of modal responses for which the model matched the data (columns "matches"). We separate these into valid and invalid syllogisms[4]. The total numbers of valid and invalid syllogisms are 24 and 40, respectively.

4.2 Discussion

5 The influence of background knowledge

The model we present in this paper is a Bayesian model, and that means the prior must be taken seriously. It's known that the content of a syllogism affects reasoning (Wilkins, 1928). This has traditionally been explored in the interaction between logical validity and the *a priori* believability of the conclusion. The effect is most prominent in the syllogistic evaluation task and has loosely been described as a tendency to endorse *a priori* believable conclusions, regardless of the logical validity of the argument. However, the degree to which conclusion-believability influences acceptance rates has been shown to be more pronounced on invalid than on valid syllogisms (J. S. B. T. Evans, Handley, & Pollard, 1983). When experimenters have included neutral materials for baseline comparisons, they find belief bias is primarily associated with *rejecting* unbelievable conclusions particularly when the argument is fallacious (i.e. an increase in correct rejections for invalids), leading some investigators to refer to it as "belief debias" (Morley, Evans, & Handley, 2004; Newstead, Pollard, Evans, & Allen, 1992).

Evans and colleagues (2001) followed up on the J. S. B. T. Evans, Handley, Harper, and Johnson-Laird (1999) study that found different endorsement rates for Possible Weak (PW) and Possible Strong (PS) syllogisms (see Section **??** for a full discussion of this finding). The investigators hypothesized that PW problems would exhibit a positive belief bias (enhancement of endorsements for believable, invalid conclusions — relative to neutral conclusions) and PS problems would exhibit

[4]Since the response format in the meta-analysis varied across studies, the number of valid syllogisms was also not the same. Here we count as valid only the syllogisms that would have been considered valid in all studies.

a negative belief bias (the typical "belief debias"). These predictions follow from the J. S. B. T. Evans et al. (1999) finding that with neutral or abstract content, PS endorsements are near ceiling and PW endorsements are near floor.

5.1 Extant theories

Much of the theoretical discussion on belief bias is concerned with *the stage of processing* where the influence of background knowledge is taken into account. According to this, theoretical accounts can be understood as arguing that background knowledge is incorporated during (1) the encoding of the problem (the translation of the syllogism into a mental representation); (2) the reasoning process (the manipulation of the mental representation); or (3) the decoding of the reasoning process (the translation of the mental representation into a response).

5.1.1 Some old qualitative theories

Selective scrutiny, misinterpreted necessity, mental models

5.1.2 Recent quantitative theories

Two recent investigations of belief bias have used computational models to disambiguate a stage 2 effect from a stage 3 effect.

Klauer, Musch, and Naumer (2000) used a Multinomial Process Tree (MPT) model to argue that that the belief bias effects occur during the reasoning process (stage 2). This comes directly from the model itself, which models the task as consisting of either (i) accurately reasoning, i.e. decided whether or not the problem is valid or invalid, and responding correctly (with probability determined by model parameter $r_{problem.type}$) or (ii) guessing (with probability $1 - r_{problem.type}$). Guessing leads to a guessing subtree, where it is assumed prior beliefs about the conclusion can have influence. Prior beliefs are modeled using parameters $\beta_{believable}$ and $\beta_{unbelievable}$. Problem type (for the reasoning parameters) consists of four possibilities, resulting from crossing believability and validity; hence, there are 4 r parameters, and these are used to test the hypothesis that the *reasoning process* is different across the 4 conditions. There are two β parameters corresponding to believable and unbelievable response biases, and these are used to test whether or not beliefs influence the response stage. The investigators found significant differences between the reasoning parameters across conditions, while constraining the β belief parameters to be equal had little effect on the fit. From this, the authors concluded belief bias is a largely an effect on the reasoning process, roughly consistent with dual-process theories of reasoning.

Dube, Rotello, and Heit (2010) drew issue with the use of Klauer et al.'s MPT models because the models assumed a "simple threshold". Dube et al. pointed out that the assumptions of threshold

models are the same as the assumptions of measuring accuracy by *Hits - False Alarms*, with which they also draw issue: that, for a given problem type and a constant level of accuracy, changes in response bias are associated with *equal changes* in acceptance rates. If the relationship between response bias and acceptance rates is empirically nonlinear however, the above assumption could lead one to infer a difference in sensitivity, when only a difference in bias is present (i.e. a Type I error on the null hypothesis that there is no difference in sensitivity / accuracy between belief conditions). To interrogate this assumption, the investigators measured ROC curves by using confidence ratings following a "valid"/"invalid" judgment. They found that the assumption of linear ROCs is unwarranted, calling into question all current models of belief bias (which implicitly use this assumption).

The authors went on to argue that the "belief bias" effect is a "response bias" effect. They do this using two null results. First, they observe that the points for believable and unbelievable ROC curves appear to lie on a single curve, "indicating subjects showed little to no difference in accuracy when judging conclusion validity". They compare the estimated area under the ROC curve for the two conditions using A_z, a SDT statistic used for accuracy when the equal-variance assumption cannot be applied[5]. They find no significant difference in the estimated area under the two ROCs. Second, they constrain the "reasoning" parameters of their Signal Detection Model ($d_{believable}$ and $d_{unbelievable}$; the distance between valid and invalid distributions of argument strength in the 2 belief conditions) to be equal. They observe a non-significant effect on the fit of the SDT models, "indicating a negligible effect of believability on accuracy".

5.2 Belief and argument strength

What can our computational-level theory tell us about the interplay between logic and belief in syllogistic reasoning?

For both the MPT and the SDT approaches, it's important to bear in mind the particular materials used to elicit the effect. As we've suggested above, the syllogistic space actually defines a distribution over argument strengths. Qualitative differences present in this distribution have been talked about before in terms of "single model" vs. "multiple model" problems (a distinction among valid syllogisms) (P. N. Johnson-Laird & Byrne, 1991) as well as "strong possible" and "weak possible" (a distinction among invalid syllogisms) (J. S. B. T. Evans et al., 1999).

The argument strength distribution depends critically on the prior distribution over situations. Though the argument strength of a valid syllogism is always 1, the argument strength of an invalid syllogism will vary with the argument as well as the content. Consider the content used in the recents studies of belief bias by Dube et al. and Klauer et al.. The content was chosen so that the subject would have strong beliefs about the truth of the conclusion (*some birds are not sparrows*)

[5]Dube et al. mention A_z is an unbiased estimator of proportion correct in a 2AFC, and that it has smaller standard error than d_a in simulations by Mamillian et al. (2004)

while remaining agnostic about the truth of the premises (e.g. by choosing an esoteric or nonsense middle term, *some birds are metazoans; no sparrows are metazoans*). We can import such a prior distribution over properties – $\Pr(bird, sparrow, metazoan)$ – into the model of argument strength to examine how such a distribution over properties shapes the prior distributions over sentences as well as the posterior distribution over argument strengths.

From the probabilistic perspective, the prior probability of *some birds are not sparrows* is very high. It's hard to know exactly what you could say to a person to convince them that *some birds are not sparrows* is not the case. This is reflected in the number of invalid syllogisms for which *some birds are not sparrows* is still highly probable [data needed].

The same argument applies for literally false (or, extremely low probability) conclusions (e.g. *some sparrows are not birds*). This is reflected in the distribution of argument strength for this content over all 64 syllogisms [plot needed]. The distribution is heavily skewed towards the end-points. The prior probability of the conclusion is so low (or so high) that there is almost nothing you could tell a person to convince them otherwise.

5.2.1 Implications for previous studies

We see that given the priors elicited from the content of previous belief bias experiments (Klauer, Dube), there is .... [probably, what we'll see is a bimodal distribution, with high information content for valid syllogisms improbable conclusions, and low informational content for everything else]. The invalid syllogisms used by Dube et al. (2010) all have very low informational content. Thus, these are relatively poor experiments for distinguishing a model of reasoning (the argument strength model) from a model of judging the conclusion only (the prior distribution over conclusions). Thus, any interpretations about the relative contribution of a "reasoning stage" based on these syllogisms should be called into question.

The strength of a valid argument is always 1, no matter the content. Thus, logically valid arguments are poor experiments for distinguishing a model that analyzes the syllogism strictly in terms of abstract content from a model that analyzes the syllogism with respect to prior knowledge. Theories of reasoning that take logical deduction as the normative theory of human reasoning have little to say about how reasoning with respect to prior knowledge might look. As such, these theories have been pushed into a strange corner of the experimental design space, and we have issue with such designs.

5.3 The current approach

Studies of belief bias in syllogistic reasoning have traditionally looked at "the conflict" of belief and logic using categorical distinctions between "believable" and "unbelievable" statements. In the most

extreme cases, "unbelievable" conclusions are taken to be those which are literally false of the real world.

We began our discussion of belief bias with the statement that the content of a syllogism can affect the conclusions drawn. In a general sense, if we consider the syllogistic sentence as a statement about two terms or properties, the properties can either be considered independent (e.g. novel words: *blickets* and *grinkiness*) or there can be some correlation between the terms (e.g. *religious people* and *church-goers*). Quantitatively, correlations can vary between -1 and +1. So far, most (possibly all?) studies of belief bias in syllogistic reasoning explore only the extreme end points of this spectrum.

Given the argument-strength distributions of the models using either necessary (*some birds are not sparrows*) or impossible (*some sparrows are not birds*) relations given prior knowledge, it would be unwise to explore potentially subtle interactions between logic and belief knowledge these materials. This is because for [most / many / all?] syllogisms, the argument strength either implies one must disregard the content (valid syllogisms) or be driven by prior beliefs (invalid syllogisms) still need to confirm these materials are no good, beyond the speculated pragmatic effects. It would be more useful to see how prior beliefs can shape reasoning when there is uncertainty in the knowledge.

6 Experiment 1

We set out to explore how the content of the syllogistic argument can affect conclusions drawn. To this end, we use content domains over which people have rich background knowledge while staying away from literally true and false propositions —the endpoints of the *believable – unbelievable* spectrum. Causal knowledge is well suited for studying belief in reasoning because (1) people have strong intuitions about causal domains and (2) causal knowledge is uncertain, thus keeping us away from the extreme endpoints of the *believable – unbelievable* spectrum [citations? Woo-kyoung Ahn?].

6.1 Hypothesis space

In this experiment, we are interested in disambiguating 3 models of reasoning. The first model is the one presented thus far: Bayesian reasoning over abstract domains (the "Abstract Bayesian" or AB). This is a model of $P(conclusion|premises)$ over situations with abstract properties. The key feature of these abstract properties is that they are determined by independent and identically distributed coin flips. This is a model of reasoning with no knowledge of the world.

The second model is a model with world knowledge but which doesn't perform any reasoning (the "World-bound Prior" or WP). This model doesn't read the premises, and responds to the conclusions based only on prior knowledge. This is a model of $P(conclusion)$ over situations generated

from world knowledge. The world-bound properties have intricate correlations between them not well modeled by i.i.d. coin flips (see Section 6.2). We determine this prior distribution over properties by asking people about the relative plausibility of these properties co-occurring (see "Prior elicitation" below).

The third model is a marriage of the previous two: Bayesian reasoning over world-bound knowledge (the "World-bound Bayesian" or WB). This is a model of $P(conclusion|premises)$ over situations generated from world knowledge. To accomplish this, we replace the independent `flips` for properties in the Abstract Bayesian with a multinomial distribution over the presence and absence of the 3 properties (again, corresponding to the 3 terms of the syllogism).

```
(define ABC (mem (lambda (x)
         (multinomial (list 'ABC 'AB_ 'A_C '_BC 'A__ '_B_ '__C '___) empirical-prior))))
```

We determined this multinomial distribution empirically, by asking Mechanical Turk participants to rate the plausibility of various combinations of properties co-occurring.

6.2 Experiment 1a — Prior elicitation

We explored the models using causal domains that fell into two structural forms: common cause and common effect.

6.2.1 Participants

We asked 70 participants on Mechanical Turk to rate the likelihood of various combinations of properties co-occurring. Participants were compensated for their work.

6.2.2 Design

We used two different dependent measures. Each participant was randomly assigned to either the "frequency" or the "plausibility" dependent measure condition. Within each of these conditions, participants completed the "frequency" or "plausibility" judgment task for all 4 domains. The design can be summarized as follows: 2 (task: "frequency" or "plausibility" judgment; between subjects) x 4 (domains: see table 3; within subjects).

6.2.3 Procedure & Materials

The most reliable way of eliciting probability judgments from subjects remains an open question. We ran the prior elicitation with two different dependent measures to examine the reliability of our materials.

The instructions for the "plausibility" condition, were as follows:

"Imagine an X (e.g. a lightbulb; see Table 3, column "Noun"). How likely is it that it:"

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 3: Content domains used in Experiment 2 syllogisms.

The instructions for the "frequency" condition were:

"Imagine 100 Xs (e.g. lightbulbs). About how many of them:"

Below these prompts were listed the 8 possible combinations of the presence and absence of the Properties A, B, C found in Table 3. In the "plausibility" condition, the properties agreed with the singular form of the noun (e.g. "is on", "is bright", and "is hot"). In the "frequency" condition, properties agreed with the plural form (e.g. "are on", "are bright", and "are hot"). All 8 combinations of the presence and absence of properties ("are on, are bright, aren't hot"; "are on, aren't bright, are hot", etc...) were listed. 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".

Participants rated all 8 combinations of properties for each domain.

6.2.4 Data analysis

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). We checked to see if these probability distributions could have been generated from independent Bernoulli random variables. We fit a 3 parameter model to each distribution independently. The correlations between the distributions generated by best fit independent Bernoulli random variables for the four domains were: X, Y, Z, and W. There is an appreciable amount of variance in the causal prior elicitation data not captured by these independent Bernoulli random variables, suggesting that we have successfully elicited distributions over prior knowledge with sophisticated correlational structure. Have I said what I wanted to say here? Do I need to say this stuff?

Syllogism selection simulations (see below) were based on the mean normalized ratings across participants.

6.2.5 Syllogism selection simulations

We are interested in disambiguating three models of reasoning: (1) the Abstract Bayesian (AB); (2) the World-bound Prior (WP) and (3) the World-bound Bayesian (WB). We used the elicited priors in WP and WB, and used a single Bernoulli random variable as the parameter `br` for AB. We compared the posterior distributions of the three models for all 64 syllogisms. We computed the expected information gain for each syllogism with the goal of disambiguating the 3 models. We choose 4 syllogisms from those with the highest expected information gain.

6.3 Experiment 1b — Syllogistic reasoning

6.3.1 Participants

We recruited 250 participants from Amazon Mechanical Turk. All participants were required to have a 95% approval rating for their previous work on the web service. Following standard practice in syllogistic reasoning experiments, we excluded subjects who had taken courses in formal logic. N participants were excluded for not listing English at their native language. M participants were excluded for having taken classes in formal logic.

6.3.2 Design

Each participant completed 4 syllogisms. Each syllogism was paired with a random domain used in the prior elicitation.

As in Experiment 1a, we used two different dependent measures to examine the reliability of our data. Each subject was either assigned to the "radio + slider" (rs) dependent measure or the "just slider" (js) dependent measure (detail in next section).

6.3.3 Procedures & Materials

The instructions to the experiment were:

> In this experiment, you will read four (4) randomly selected logical arguments. For each argument, you will be presented with different conclusion that might follow from the argument presented.

On each experimental slide, the subjects saw the words "Given that:" followed by the syllogism. Below was written: "Does it follow that:". Below that was presented a 4 column table with each of the 4 syllogistic conclusions in a column. Below each conclusion was the dependent measure.

In the rs condition, the dependent measure was a radio button with the options "Doesn't follow" and "Follows". Below that was a vertically-oriented slider bar with endpoints labeled "Certain"

and “Don’t know”. In the js condition, the dependent measure was just a slider bar with endpoints labeled “Certainly follows” and “Certainly does not follow”.

Below this, in either condition, was listed the following instruction:

> If you think the conclusion follows from the argument, indicate so on the line. Adjust the position of the slider bar to reflect your confidence in your response.

Participants were required to mark each slider bar before continuing on to the next page.

6.3.4 Analysis

The rs dependent measure responses were transformed onto a 0–1 scale by saying that the “Don’t know” slider value corresponds to a 0.5 response and that any deviation from that is reflected in either a positive or negative way determined by the radio button. The responses for each subject for each syllogism were normalized so they made a well-formed probability distribution (i.e. they added up to 1).

The correlation between the rs responses and the js responses was VERY HIGH. Thus, we collapsed across the two dependent measure condition.

There were effects of content and of syllogism on conclusions drawn. [import boring frequentists statistics here: ANOVAs? Chi-squares?]

6.3.5 Model fits

The World-bound Bayesian model has 1 parameter: the number of objects in a situation. The World-bound Prior has the same single parameter. The Abstract Bayesian model has 2 parameters: the number of objects and the base rate of properties. We fit these model parameters to the data to maximize the correlation between the model and the data.

Modal response hits

Correlations

6.3.6 Bayesian model comparison

6.4 Discussion

7 Experiment 2

We’ve seen so far that the content of the syllogism affects reasoning in a way perfectly compatible with the Bayesian notion of a prior distribution using our probabilistic model specification. The World-bound Bayesian model captures the modal responses as well as the graded quantitative data inherent in subjects’ responses to these syllogisms. The models so far, however, use the standard

semantics of the quantifiers and as such, are unable to show a preference among equally valid conclusions. This is apparent in Aristotle's so-called *perfect syllogism*:

All As are Bs

All Bs are Cs

This is one of the easiest syllogisms insofar as most people draw a correct valid conclusion (Khemlani & Johnson-Laird, 2012). The valid conclusion they draw is *All As are Cs*. There is another valid conclusion however, namely *Some As are Cs*. Most respondents do not seem to recognize this as a valid alternative. If people were following the standard logical meaning of the quantifiers, we would expect subjects to draw each valid conclusion about 50% of the time. Subjects do not do express this ambivalence, however. They greatly prefer the *All* conclusion. This is apparent even when the two complete arguments are presented separately in the Evaluation task (?, ?).

8 Pragmatics in syllogistic reasoning

It's been suggested that the reason for this asymmetry among equally valid conclusions is a result of a scalar implicature cite somebody. Scalar implicature is the phenomenon where a relatively weak utterance is thought to imply the negation of a stronger alternative. The most well studied of these scales is the quantifier scale, which is thought to consist of none, some, and all, in increasing order.

8.1 Semantics and presuppositions

In the model presented so far, we have assumed all properties are instantiated at least once in every situation, what linguists and philosophers call *existential import*. We will call this semantics where everything has existential import as "plentify world semantics". This idea is that quantifier sentences cannot be vacuously true (e.g. *All As are Bs* cannot be true if there are no *As*).

There are other reasonable assumptions about the semantics of the quantifier sentences that are worth considering. A somewhat more permissive set of definitions would consider that the universal quantifiers *none* and *all* could be vacuously true cite Aristotle. We might call this the "smart aleck semantics", following in the steps of Guerts, presupposition paper reference. Under this semantics, the particular quantifier *some* would still include the existential claim that "there exists at least one".

In between the "smart aleck semantics" and the "plentify world semantics" is one in which you take Aristotle's point of view (the SAS view) but claim that the definite determiner phrase "of the" grants existential import to the set in question (i.e. "All blickets are red" doesn't necessarily imply that there are any blickets at all but "All of the blickets are red" means that there are some blickets [and hence, also red things]). We will call this view the "definite determined semantics" or DDS.

This in turn means that the subject term of each syllogistic sentence has existential import. This semantics will still allow empty terms for the quantifiers "none" and "some...not".

We leave for future work a more fine-grained test of the psychological reality of these alternative semantics in the syllogistic reasoning task. Instead we briefly demonstrate that at the literal level, all semantics given roughly similar results, though at the pragmatic SAS and DDS provide somewhat counterintuitive and implausible results.

OED on these different models? probably just to show some examples

8.2 Alternatives and pragmatic inference

It's known that the salience of different alternatives of what a speaker could have greatly impacts the quality and strength of a pragmatic inference. In the pragmatics syllogism model, we considered the abstract alternative space of all possible other syllogistic premises (or, the 15 syllogistic premises of the same figure) as the space of things the experimenter could have produced. Though this is consistent with the generative model of many syllogism experiments (syllogisms sampled uniformly from the space of all possible syllogisms), there are other reasonable ways of specifying the alternative set.

1. Single-edit distance on quantifier
2. Single-edit distance of quantifier / term -ordering
3. Syllogisms of the same figure
4. All possible syllogisms

9 Formal semantics and generalized quantifiers

10 General discussion

10.1 Can nothing ever actually follow?

10.2 On the meaning of *deduction* and epistemic modals

Many studies of syllogistic reasoning run under the assumption that there are two modes of reasoning: formal and everyday. Task instructions for these studies often stress to draw the conclusion "that can be deduced with *absolute* certainty" or "that follows *necessarily* from the premises"(Khemlani & Johnson-Laird, 2012).

Recent work on epistemic modal words like *possible*, *probable*, and *necessary* (Lassiter & Goodman, 2013) suggests that differences in reasoning elicited by these words can be accounted for by a one-dimensional model of reasoning.

It should also be noted that the word *deduction* has both a technical meaning and a lay meaning. In the logician's book, deduction refers to reasoning from a rule to a particular instance. If the rule is true, and the terms are clear, then the conclusion is necessarily true. In this way, deductive reasoning is infallible. The lay meaning of the word *deduction*, however, is not so clear. One very popular source of the word is the fictional character Sherlock Holmes. Many of Sherlock Holmes' "deductions", however, are considered by scholars to be creative forms of abduction (Eco, 1983). The lay meaning of *deduction* may be more similar to reasoning or generally, inference.

10.3 The prior distribution over conclusions

Since we now have a generative model of the argument-strength of a syllogism — $P(\text{conclusion} \mid \text{premises})$, it makes sense to think about the prior probability of a conclusion: $P(\text{conclusion})$. Since the `(sample-conclusion)` function performs a `uniform-draw` over conclusions, the model has maximal uncertainty about the conclusion *before situations are sampled*. Consider this as a traditional multiple choice test where the test-taker has no bias towards *A, B, C,* or *D*, as such. For each situation sampled, however, some conclusions will be true of it and others will be false of it. After the distribution of situations has been explored, there will be a distribution over true conclusions. This is the prior distributions over conclusions. Because situations are sampled according to a prior distribution over properties (in the most basic case above, 3 independent `flips`), the corresponding prior of conclusions is unlikely to be uniform.

10.4 Not all invalid syllogisms are weak arguments

Evans and colleagues (1999) had participants rate all possible combinations of 64 syllogistic premises and 4 conclusions using an evaluation task with abstract content (letters e.g. D, H, Z for terms). The study was aimed at examining the differences between instructions ("what is a {*necessary, possible*} conclusion?") to test some predictions of the mental models framework.

The investigators discovered unexpectedly that among syllogisms with possible (but not necessary) conclusions (i.e. *fallacies*, in the deductive sense), some are very strongly endorsed (in both instructional conditions) while others are very weakly endorsed. They replicated this finding in a separate experiment, in which participants were presented with four problem types: Necessary (i.e. valid), Possible Strong (PS), Possible Weak (PW), and Impossible (a contradiction of a Necessary conclusion). They found that the PS problems were endorsed about as often as the Necessary problems, and the PW problems were endorsed almost as little as Impossible problems.

The notion of Possible Strong and Possible Weak syllogisms falls right out of analyzing the argument strength of syllogisms.

[Plot here a histogram/density-plot of argument strength of 256 sylls X 2 term orderings and

highlight the possibly weak and possible strong ones (i.e. from the Evans study).]

10.5 The informational content of a syllogism

Evans and colleagues (1999) found that some invalid syllogisms are endorsed more strongly than other invalids. This suggests that some invalid syllogisms are actually relatively strong arguments. To put this in quantitative terms, we can analyze the informational content of the syllogism, as defined by the degree to which the syllogism updates our prior beliefs over sentences into our posterior beliefs — the argument strength distribution. We can formalize this by using the expected KL divergence between a prior distribution over conclusions and a posterior distribution of argument-strength.

What we find is roughly an exponential distribution of informational content of syllogisms [insert exponential distribution of informational content of syllogisms here]. Many of the most informative syllogisms are valid syllogisms; this is because valid syllogisms are maximally strong and these arguments lead to conclusions very different from the prior over conclusions.

The informational content of a syllogism will depend on the prior over conclusions, which itself depends upon the prior distribution over conclusions. This will be important to bear in mind when we revisit these considerations in the next section: the influence of background knowledge.

References

Dube, C., Rotello, C. M., & Heit, E. (2010, July). Assessing the belief bias effect with ROCs: it's a response bias effect. *Psychological review*, *117*(3), 831–63. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/20658855` doi: 10.1037/a0019634

Eco, U. (1983). The sign of three: Dupin, holmes, peirce. In U. Eco & T. A. Sebeok (Eds.), (p. 198-220). Bloomington: Indiana University Press.

Evans, J. S., Handley, S. J., & Harper, C. N. (2001, August). Necessity, possibility and belief: a study of syllogistic reasoning. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, *54*(3), 935–58. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/11548042` doi: 10.1080/713755983

Evans, J. S. B. T., Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning About Necessity and Possibility : A Test of the Mental Model Theory of Deduction. *Journal of experimental psychology. Learning, memory, and cognition*, *25*(6), 1495–1513.

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.

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church : a language for generative models. *Uncertainty in Artificial Intelligence*.

Johnson-Laird, P. (1983). Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction* (L. Erlbaum, Ed.). UK: Hove.

Johnson-Laird, P. N., & Steedman, M. (1978). The Psychology of Syllogisms. *Cognitive psychology*, *10*, 64–99.

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. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.107.4.852 doi: 10.1037//0033-295X.107.4.852

Lagerlund, H. (2012). Medieval theories of the syllogism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2012 ed.). http://plato.stanford.edu/archives/win2012/entries/medieval-syllogism/.

Lassiter, D., & Goodman, N. D. (2013). How many kinds of reasoning? Inference, probability, and natural language semantics. *Cognitive science*.

Lassiter, D., & Goodman, N. D. (2014). How many kinds of reasoning ? Inference , probability , and natural language semantics. *submitted*(2014), 1–22.

Morley, N. J., Evans, J. S. B. T., & Handley, S. J. (2004, May). Belief bias and figural bias in syllogistic reasoning. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, *57*(4), 666–92. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15204128 doi: 10.1080/02724980343000440

Newstead, S. E., Pollard, P., Evans, J. S., & Allen, J. L. (1992, December). The source of belief bias effects in syllogistic reasoning. *Cognition*, *45*(3), 257–84. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1490324

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

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: the probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Störring, G. (1908). Experimentelle untersuchungen uber einfache schlussprozesse. *Arch. f. d. ges. Psychol*, 1-127.

Wilkins, M. C. (1928). *The Effect of Changed Material on Ability to do Formal Syllogistic Reasoning*.