Some arguments are probably valid: Syllogistic reasoning as communication

Michael Tessler, Noah D. Goodman
{mtessler, ngoodman}@stanford.edu
Department of Psychology, Stanford University

Abstract

Syllogistic reasoning lies at the intriguing intersection of natural and formal reasoning, of language and logic. Syllogisms comprise a formal system of reasoning yet use natural language quantifiers, and invite natural language conclusions. How can we make sense of the interplay between logic and language? We develop a computational-level theory that considers reasoning over concrete situations and situations to be constructed probabilistically by sampling. The base model is enriched to consider the pragmatics of natural language arguments. The model predictions are compared with behavioral data from a recent meta-analysis. The flexibility of the model is explored in a published study of syllogisms using the generalized quantifiers *most* and *few*. We conclude by relating our model to two extant theories of syllogistic reasoning – Mental Models and Probability Heuristics.

Keywords: Reasoning; language; Bayesian model

Consider for a moment that your friend tells you: "Everyone in my office has the flu and, you know, some people with this flu are out for weeks." Do you respond with "Everyone in your office has the flu." Do you respond with "Pardon me, there is no inference I can draw from what you just said." Or do you respond "I hope your officemates are not out for weeks and I hope you don't get sick either."

The first response – while true – does not go beyond the premises; the second response attempts to go beyond the premises by strict classical logic, and fails; the final response goes beyond the premises, to offer a conclusion which is probably useful and probably true. This cartoon illustrates a critical dimension along which cognitive theories of reasoning differ: whether the core and ideal of reasoning is deductive validity or probabilistic support. A separate dimension concerns the extent to which principles of natural language—pragmatics and semantics—are necessary for understanding reasoning tasks. In this paper we explore a theory of syllogistic reasoning inspired by recent advances in probabilistic semantics and pragmatics.

The form of the argument above resembles a syllogism: a two-sentence argument used relate two properties (or terms) via a middle term; the relations used in syllogisms are quantifiers. Fit into a formal syllogistic form, this argument would read:

All officemates are out with the flu
Some out with the flu are out for weeks
Therefore, some officemates are out for weeks

The full space of syllogistic arguments is derived by shuffling the ordering of the terms in a sentence ("All A are B" vs. "All B are A") and changing the quantifier (*all, some, no, some ... not*). Most syllogisms have no valid conclusion, i.e. there is no relation between *A* & *C* which is true in every situation in which the premises are true. This is the case with

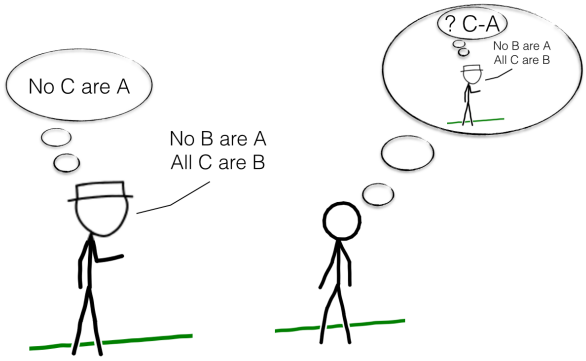

Figure 1: How will the reasoner interpret the experimenter's argument?

the argument above. Often in these cases, however, people are perfectly comfortable drawing some conclusion. A recent meta-analysis of syllogistic reasoning showed that over the population, the proper production of *no valid conclusion* responses for these invalid arguments ranged from 76% to 12%. For valid arguments, the accuracy of producing valid conclusions ranged from 90 % to 1% (Khemlani & Johnson-Laird, 2012): people do not seem to find drawing deductively valid conclusions particularly natural.

Perhaps because of this divergence between human behavior and deductive logic, syllogistic reasoning has been a topic of interest in cognitive psychology for over a hundred years (Störring, 1908), and before that in philosophy, dating back to Aristotle. Syllogisms are undoubtedly logical; indeed, the syllogistic form was the only formal system of logic for millennia. At the same time, the arguments use natural language quantifiers and invite natural language conclusions; precisely pinning down the meaning and use of quantifiers has been an ongoing area of inquiry since Aristotle (e.g. Horn, 1989).

Many theories of syllogistic reasoning take deduction as a given and try to explain errors as a matter of noise during cognition. Errors, then, may arise from improper use of deductive rules or biased construction of logical models. Many other kinds of reasoning, however, have been explained as probabilistic inference under uncertainty. Probability theory provides a natural description of a world in which you don't know exactly how many people are in the hallway outside your door, or whether or not the lion is going to charge. We suggest that combining probabilistic reasoning with natural language semantics and pragmatics is a useful approach: an approach in which knowledge describes distributions on pos-

sible situations, and in which these distributions can be updated by sentences with new information. In this formalism, deduction emerges as those arguments which are always true and syllogistic reasoning becomes a process of determining that which is most probable, relevant, and informative.

A Bayesian reasoner model

Our model begins with the intuition that people reason probabilistically about situations populated by objects with properties. To represent this type of richly structured model, we must go beyond propositional logic and its probabilistic counterpart, Bayesian networks. We instead build our model using the probabilistic programing language Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008), a kind of higher-order probabilistic logic in which it is natural to describe distributions over objects and their properties. For background and details on this form of model representation, see `http://probmods.org`.

Situations are composed of *n* objects:

```
(define objects (list 'o1 'o2 ... 'on))
```

(Ellipses indicate omissions for brevity, otherwise models are specified via runnable Church code.) Properties `A`, `B`, and `C` of these objects are represented by functions from objects to the property value. We assume properties are Boolean, and so property values can be `true` or `false`. We assume no *a priori* information about the meaning of the properties and 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))))
```

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 determined probabilistically (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[1]. For us, this simply means the base rate, `br`, of properties is small.

We interpret quantifiers as truth-functional operators, consistent with standard practice in formal semantics. A quantifier (e.g. *all*) is a function of two properties (e.g. *As* and *Bs*) 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`, `no`, `some-not` to have their standard meanings. For a first test of the model, we assume sets are non-empty, i.e. *all* and *none* cannot be trivially true.

The key observation to connect these 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, these quantifier expressions 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:

```
(query
 (define objects (list 'o1 'o2 ... 'on))
 ...define A,B,C...
 ...define all, some, no, some-not...
 (define conclusion (conclusion-prior))

 conclusion

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

The first arguments to a query function are a generative model: definitions or the background knowledge with which a reasoning agent is endowed. Definitions with which a prior is stipulated (e.g. `conclusion`) denote aspects of the world over which the agent has uncertainty[2]. 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.

We assume that the prior distribution over conclusions (and premises, below) is uniform.

Recursion and pragmatics

We have suggested viewing syllogistic reasoning as a case of communication, and this in turn suggests that reasoning should go beyond the semantics of language, to its pragmatics.

Following the *rational speech-act* (RSA) theory (Goodman & Stuhlmüller, 2013; Frank & Goodman, 2012), we imagine a reasoner who receives premises from an informative speaker. The speaker conveys information about which only she has access – in RSA, her access was a current state of the world. By being informative with respect to the world-state, the speaker is able to communicate enriched meanings (e.g. scalar implicature – that "some" may also imply "not all"). It is known, however, that standalone scalar implicatures do a poor job of accounting for reasoning with syllogisms (M. J. Roberts, Newstead, & Griggs, 2001). Indeed, a preliminary analysis of a standard Gricean-listener model in this framework was consistent with this account.

However, a listener (our reasoner) may consider the premises in a wider, conversational setting: she may ask herself why the experimenter chose to give these particular

[1]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.

[2]Note that for brevity, the properties `A, B, C` are wrapped-up; together with objects, these properties generate situations, and this is also an aspect over which the agent has uncertainty.

premises, as opposed to alternative arguments. This requires a closer look at what the reasoner believes to be at issue in this "conversation"—the Question Under Discussion, or QUD (C. Roberts, 2004). In a syllogistic context, we take the QUD to be "what is the relationship between A & C (the end terms)?", very often the actual context in which the experiment is presented.

In this setup, pragmatic inferences will differ from the standard local implicatures; for instance, "Some A are B" may not lead to a "Not all A are B" implicature if "All A are B" wouldn't provide additional information about the A-C relationship. The enriched meanings come from the following counter-factual consideration: "why did this experimenter present me with this argument and not any other argument?" The pragmatic reasoner enriches the conclusions that are more uniquely determined by the particular argument the experimenter provides.

The A-C QUD is naturally captured by a `reasoner` who considers an `experimenter` who considers the conclusion the `reasoner` would draw about A & C (not the reasoner's inferences about the whole world-state, which would — superfluously — include B).

We can combine the above intuitions about pragmatic comprehension into a model in which `reasoner` and `experimenter` jointly reason about each other. Critically, each agent reasons about the other at recursive depth `depth` of comprehension:

```
(define (experimenter conclusion depth)
  (query
   (define premises (premise-prior))

   premises

   (equal? conclusion (softmax (reasoner premises depth)
        alpha))))

(define (reasoner premises depth)
  (query
   (define objects (list 'o1 'o2 ... 'on))
   . . . define A,B,C . . .
   . . . define all, some, no, some-not . . .
   (define conclusion (conclusion-prior))

   conclusion

   (and (conclusion A C)
     (if (depth 0)
          (and ((first premises) A B)
               ((second premises) B C))
          (equal? premises (experimenter conclusion (-
               depth 1)))))))
```

The `reasoner` and `experimenter` functions produce a distribution over conclusions and premises[3], respectively. Since we take these functions to represent actual persons in a communicative setting, we take premises to be selected from these distributions according to a Luce choice, or `softmax`, decision rule with a parameter `alpha` that denotes the degree to which argument is chosen optimally (Luce, 1959). This takes the distribution, raises it to a power `alpha` and renormalizes. As `depth` increases, the premises becomes more informative with respect to the uniquely-implicated conclusions (for those premises); i.e. the arguments are interpreted as more tailored for a particular conclusion. When *depth* is 0, the model collapses to produce the $P(\text{conclusion} \mid \text{premises})$, which we refer to as the *literal bayesian reasoner*. We refer to the model with *depth* greater than 0 as the *pragmatic bayesian reasoner*.

[3]As a first pass, we consider the alternative premises generated by `premise-prior` to be the set of all premises of the same term orderings, i.e. all premises of the same *figure*. That is, we consider alternative quantifiers, keeping the structure of the sentence fixed.

Results

To test the predictions of the model we used data from the meta-analysis of syllogistic reasoning tasks presented by Chater and Oaksford (1999). These data were compiled from five studies on syllogistic reasoning, completed from 1978-1984. The data include 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[4]. We removed the NVC responses from the meta-analysis and renormalized so the endorsement of all conclusions for a syllogism adds to 1. 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.

The three parameters of the model — `n_objects`, `br`, and `alpha` — were fit via grid-search to produce the maximum likelihood of the data given the model. These fit parameter values were 5, 0.25, and 4.75, respectively[5].

Qualitative 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[6]. The total numbers of valid and invalid syllogisms are 24 and 40, respectively.

Model	matches_{valid}	$\text{matches}_{invalid}$	r_{valid}	$r_{invalid}$
Prior	5	24	-.46	.41
Literal	17	20	-.20	.64
Pragmatic	17	26	.77	.74

As a baseline, we first examined the posterior distribution of conclusions conditioned only on the truth of the conclusion (what we refer to as the "Prior") to see if it alone accounted for human reasoning patterns. It did not (Figure 2, column

[4]This is because in each possible situation, one of the four conclusions will be true. In fact, since the four possible conclusions form two pairs of logical contradictions, exactly two conclusions will be true in each situation.

[5]N.B. `n_objects` = 5 keeps the number of distinct, possible objects in a situation to a minimum while `br` = 0.25 is in accord with the rarity assumption.

[6]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.

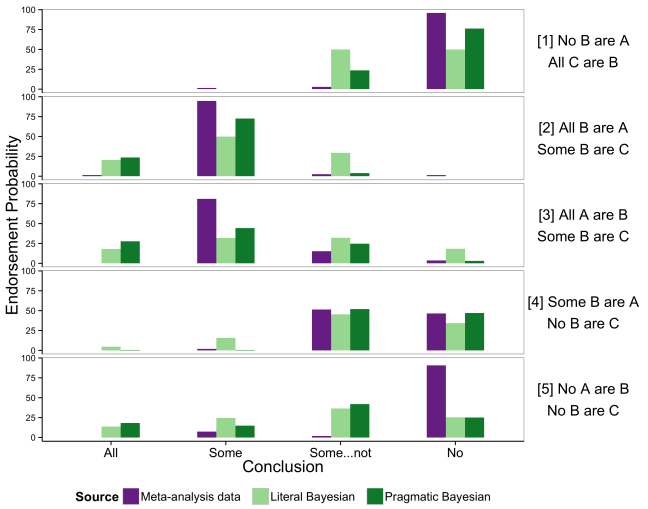

Figure 2: Five example syllogisms. [1] Literal reasoner has no preference among equally valid conclusions; the symmetry is broken by the pragmatics models which considers the argument in the space of possible arguments. [2] Literal reasoner alone captures the modal response and pragmatics enriches the quantitative fit. [3] Relatively uninformative premises suggest *some* is the most likely interpretation. [4] Models are able to capture multiple preferred conclusions. [5] Models do poorly in matching subjects' responses in an underconstrained, invalid syllogism.

1). Since *some X are not Y* is the most likely conclusion to be true, the Prior matches only the syllogisms with a *some...not* modal response. The literal Bayesian reasoner matches the modal response on 37 of the 64 syllogisms. The 29 syllogisms for which *some...not* was the modal response are qualitatively unaffected. The model also matches 8 syllogisms for which *some* and *none* are favored (Figure 2, column 2, see e.g. [2]). Probabilistic reasoning introduces gradation which accounts for an appreciable portion of the variance.

Conversational pragmatics can enrich the meaning of the premises given to the reasoner — PBR — by considering "why has the experimenter produced this argument — these premises — given that she may have given other arguments?" PBR maximally-prefers the modal response of subjects for 43 out of 64 syllogisms. As well, it picks up on some of the very complex phenomena present in syllogistic reasoning. Example [3] in Figure 2 is one such case. The premises considered literally are relatively uninformative. The literal reasoner is very similar to the Prior (not shown in Figure 2; but see Figure, 3 column 1).

In addition to capturing many of the modal responses, the model is able to accommodate more than one plausible conclusion. [4] in Figure 2 is one such example. This is a syllogism with a valid conclusion, but one which people find difficult to draw. The literal reasoner model tells us why: in many of the possible situations in which the premises are true, a *none* conclusion is true. In addition, *none* is more difficult to convey in an argument—relative to *some...not*—and so the pragmatic Bayesian strengthens the plausible but invalid *none*.

PBR considers the space of all arguments. Most arguments in the syllogistic space are relatively vacuous, i.e. they do not update the prior substantially. As such, the most probable conclusion given all arguments is *some X are not Y*. Since the argument in [3] is more informative relative to others (e.g. the argument in [5]), the most likely intention of the experimenter is to convey that *some A are C*.

Though this is encouraging qualitative data, there are a number of syllogisms for which reasoning patterns are not accounted for by PBR. Many of these are syllogisms use two negative quantifiers (*some...not* or *none*) as the premises. For these arguments, the predictions of the literal reasoner do not differ appreciably from the predictions of the Prior (Figure 2 [5]), because the rarity prior assumes most relations will be false to begin with.

Model fit

To assess our models' quantitative fits we examine correlations across all 256 data points (64 syllogisms x 4 conclusions), shown in Figure 3. The Prior's predictions are the same for all syllogisms and the overall fit is poor (r = 0.36). After conditioning on the truth of the premises, the model is able to make graded responses. These responses are a reflection of the types of situations consistent with the premises. The overall correlation is appreciably higher (r = 0.64). Among valid conclusions, however, (squares in Figure 3) the fit is terrible (r = -0.20 for valid conclusions only). This is a direct consequence of the reasoner's literalness: the model has no preference among multiple valid conclusions, since a valid conclusion – by definition – is one which is true in every situation in which the premises are true.

This symmetry is broken by the reasoner who interprets the premises as coming from a pragmatic experimenter (Figure 3, column 3), and the overall fit improves (r = 0.77). The model is now able to make graded responses among valid conclusions (r = 0.77 for valid conclusions only).

Model flexibility: generalized quantifiers

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. The meaning of generalized quantifiers like "most" and "few" is a topic of debate in formal semantics, but can be modeled to a first approximation as a thresholded function. As a first test of the generality of the model, we define most and few by a threshold of 0.5 such that "most As are Bs" is true if more than half of the As are Bs. Once we have added these lexical items, the bayesian reasoning models extend naturally. We compare our model predictions to two studies carried out by Chater and Oaksford (1999) on syllogisms using the generalized quantifiers *most* and *few* e.g. *Most artists are beekeepers; Few chemists are beekeepers*. Participants were told to indicate which, if any, of the four quantifier conclusions followed from the premises

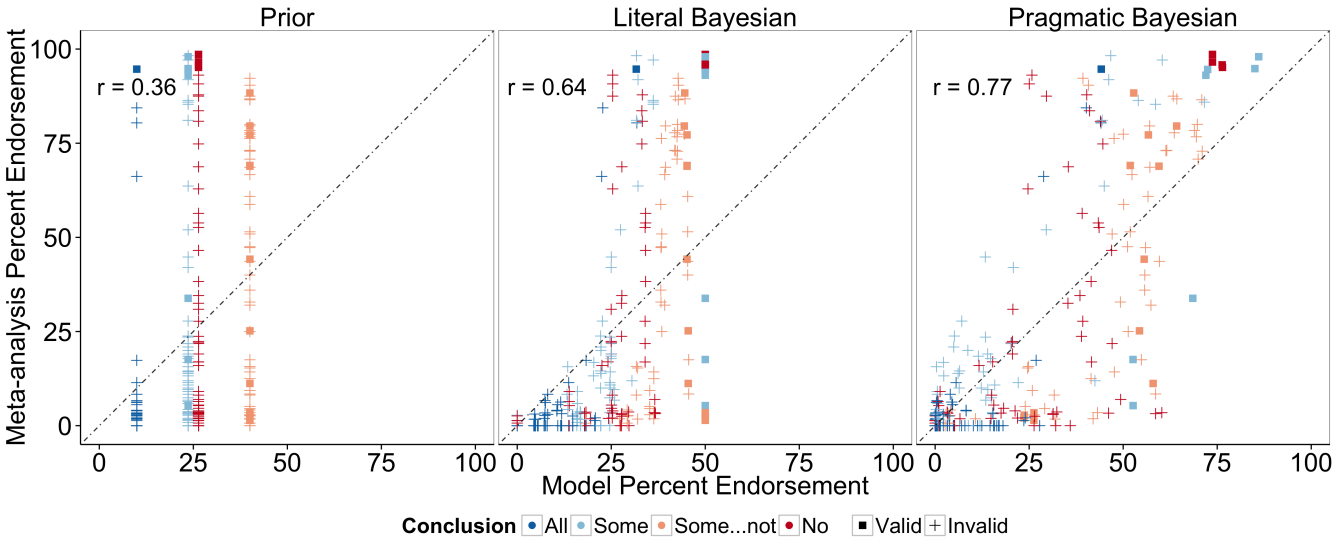

Figure 3: Human subject percentage endorsement vs. model predictions. Columns (from L to R): predictions based only on the prior [P(conclusion)]; literal Bayesian reasoner [P(conclusion | premises)]; and the pragmatic Bayesian reasoner (see text).

and were allowed to select multiple options. The set of syllogisms was divided into two experiments to avoid subject fatigue.

We find good correspondence between the experimental data and the model, even with only a local parameter search[7] (Figure 4). In Experiment 1, the quantifiers *all*, *most*, *few*, and *some...not* were used. In Experiment 2, the quantifiers *most*, *few*, *some*, and *none* were used. Note again the total number of syllogisms in an experiment is 64.

Model	matches$_{Exp_1}$	matches$_{Exp_2}$	r$_{Exp_1}$	r$_{Exp_2}$
Prior	23	23	.55	.34
Literal	42	36	.79	.65
Pragmatic	47	35	.83	.67

The fit is appreciably better for Experiment 1 than for Experiment 2, and the same was true for the Probability Heuristics Model (r = 0.94 vs r = 0.63). Overall, the proportion of *no valid conclusion* responses in the experimental data, which we do not model, was much higher in Experiment 2 than in Experiment 1. This may explain why the pragmatic reasoner tends to give high endorsement to many conclusions which people do not. A model that takes into account NVC may alleviate this effect.

Discussion

The inspiration for the pragmatic bayesian reasoning model comes from the idea that syllogistic reasoning cannot be disentangled from language understanding. Natural language semantics alone seems to be insufficient to explain the variability in reasoning, however. We have shown that a combination of semantics and conversational pragmatics provides insight into how people reason with syllogistic arguments.

A recent meta-analysis carved the space of reasoning theories into three partitions: those based on models or diagrammatic reasoning, those based on formal logical rules, and those based on heuristics (Khemlani & Johnson-Laird, 2012). We see the space slightly differently. In one dimension, theories may be based on the direct application of derivation rules—be they heuristic or logical—or they may be based on the construction of concrete representations or models. In another dimension, theories may fundamentally be interested in deductive validity or probabilistic support. This theoretical partitioning places the Bayesian reasoning models in a previously unexplored quadrant of the two-dimensional theoretical space described: we consider probabilistic reasoning over concrete situations.

Mental Models Theory (MMT) was offered to capture the intuition that people are able to reason about sets of things explicitly and with respect to context by constructing mental representations of individuals over which to reason. The situations described in our computational models are analogous to mental models. MMT, however, leaves the problem of determining which models come into existence to complex heuristics. By contrast, we derive a distribution over models or situations from natural language semantics and pragmatics, with no further assumptions.

Chater and Oaksford (1999) introduced the Probability Heuristic Model (PHM) which derives a set of probabilistic rules for syllogistic reasoning; to account for informativity and other effects, the PHM then augments these probabilistic rules with a complex set of heuristics (for example, informative-conclusion heuristics). Our model differs

[7] `n_objects` fit to 6, `br` to 0.30, `alpha` to 4.75. The words "most" and "few" might pragmatically implicate sets of substantially larger size, and thus the data might be captured better by searching over a larger parameter space for `n_objects`. In this analysis, we examined only a small search radius around the parameter estimates used to model the meta-analysis data.

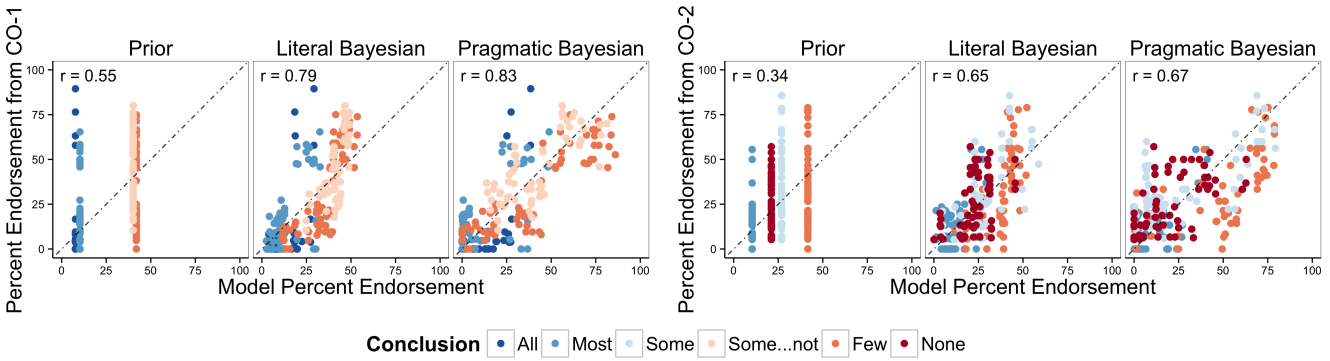

Figure 4: Human subject percentage endorsement vs. model fits for 2 experiments using generalized quantifiers. Experiment 1 (left) used the quantifiers {*all*, *most*, *few*, *some...not*}. Experiment 2 (right) used the quantifiers {*most*, *few*, *some*, *none*}.

in two respects. First, the probabilistic "rules" emerge naturally from the semantics of quantifiers by reasoning about situations. Second, we strengthen inferences by employing previously-proposed formalisms for pragmatic reasoning. This gives rise to many of the same effects, such as informativity, without postulating heuristics *de novo*.

The syllogistic reasoning task involves reading a pair of sentences and producing or evaluating a conclusion. We have considered the pragmatics of argument interpretation — the problem the reasoner faces when given some sentences. Natural language pragmatics may also enter into the production of a conclusion (for tasks that require production). The reasoner is likely tempted to produce conclusions which are not only true but also *good*, or informative. At the same time, the option of "no valid conclusion" – of saying nothing – looms large for the reasoner. We leave for future work the incorporation of production of informative conclusions as well as the ability to say "nothing follows".

Conclusion

This is early work and we have found promising evidence, both qualitative and quantitative, that this framework will allow for a more explicit understanding of syllogistic reasoning.

A major virtue of the pragmatic reasoning framework is that it extends naturally to incorporate any terms for which a truth-functional semantics can be given. For instance, we tested the model on *most* and *few* using the simplest, most standard semantics (most is more than half, etc). It is likely that these quantifiers actually have more complex semantics, but even so we accounted for a significant fraction of the data.

In this framework, a syllogism is read as an argument given as a part of discourse between interlocutors. Indeed, this is how syllogisms were used in the time of Aristotle and in the long tradition of scholastic philosophers since. Fundamentally, syllogisms are a tool used to convince others. The results of the Pragmatic Bayesian Reasoner recast the ancient idea that human reasoning behavior is as much reason as it is human. Gauging degrees of truth or plausibility alone is not sufficient. An agent needs to be posited at the other end of the line so that a conclusion makes sense; so that an argument may convince!

References

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

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

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*.

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

Horn, L. R. (1989). *A natural history of negation*. Chicago: University of Chicago.

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

Luce, R. D. (1959). *Individual choice behavior*. New York, NY: Wiley.

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

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT Press.

Roberts, C. (2004). Information structure in discourse. *Semanatics and Pramatics*(5), 1-69.

Roberts, M. J., Newstead, S. E., & Griggs, R. a. (2001, May). Quantifier interpretation and syllogistic reasoning. *Thinking & Reasoning*, *7*(2), 173–204.

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