Some arguments are probably valid: Reason and language in syllogisms

M. H. Tessler, Noah D. Goodman
{mhtessler, ngoodman}@stanford.edu
Department of Psychology, Stanford University

Abstract

We develop a computational-level theory of syllogistic reasoning which places reasoning at the intersection of communication and logic. We compare our model predictions with behavioral data from a recent meta-analysis. We show the flexibility of the model to account for reasoning behavior in a study of so-called "statistical syllogisms" which use the generalized quantifiers *most* and *few*. We relate our model to three extant theories of syllogistic reasoning – Mental Models, Mental Logics and Probability Heuristics. We conclude by discussing further predictions of the model and future directions.
Keywords: Reasoning, language understanding, probabilistic model

Consider for a moment that your friend tells you: "Everyone in my office has the flu and, you know, some people with the 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 is true, but does not go beyond the premises; the second response attempts to go beyond the premises by strict classical logic, but fails; the final response goes beyond the premises, to offer a conclusion which is probably useful and probably true. This cartoon illustrates two dimensions along which cognitive theories of reasoning differ: whether the core and ideal of reasoning is deductive validity or probabilistic support, and, the natural language principles—pragmatics and semantics—for understanding reasoning. 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: an argument with two quantifier expressions (premises) used to relate two properties (or terms) via a middle term. Fit into a formal syllogistic form, this argument would read:

All officemates are out with the flu
Some people out with the flu are out for weeks

The full space of syllogistic arguments is derived by shuffling the term-ordering ("All A are B" vs. "All B are A") and changing the quantifier (*all, some, none, not all*). Most syllogisms have no valid conclusion, i.e. there is no deductive relation between A & C determined by the premises. This is the case with the argument above. A recent meta-analysis of syllogistic reasoning tasks showed that over the population, accuracy on producing valid conclusions ranges from 90 % to 1% and ability to produce appropriate *no valid conclusion* responses ranges from 76% to 12% (Khemlani & Johnson-Laird, 2012): people are not good at drawing deductively valid conclusions. Perhaps because of this décolage between human behavior and deductive logic, syllogistic reasoning has been a topic of considerable interest in cognitive psychology for over one hundred years (Störring, 1908), and before that in philosophy, dating back to Aristotle. In cognitive psychology we are interested in how people reason, and syllogisms lie at the intriguing intersection of natural and formal reasoning, of language and logic. They are undoubtedly logical; indeed, they are regarded as the first formal system of logic. At the same time, they 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 errors in the system. These error may arise from improper use of deductive rules, or biased construction of logical models. In either case, logic makes the connection to natural language semantics natural (though not central). On the other hand, many other kinds of reasoning have been explained via probabilistic reasoning under uncertainty. Probability provides a natural description of a world in which you don't know how many people are in the hallway outside your door or whether or not the lion is going to start charging. We will suggest that combining probabilistic reasoning with formal semantics and pragmatics of natural language leads to a useful combination of these approaches, in which our knowledge describes distributions on possible situations and sentences naturally update these distributions with new information. In this formalism, deduction emerges as those arguments which are always true and syllogistic reasoning becomes a process of determining what is most probable, relevant, and informative.

The Conditional Semantics model

The motivation for our model stems from the intuitions that people reason by constructing situations consistent with their probabilistic prior knowledge, and they interpret premises and choosing conclusions as

: (1) people are using *everyday* reasoning in syllogistic reasoning tasks and (2) they do this by constructing situations and reasoning about these situations. For syllogistic reasoning, it is natural to think of situations as composed of objects with properties[1]. In a probabilistic model, situations arise from a sampling procedure.

Situations are composed of n number of objects, each with 1, 2 or 3 properties corresponding to the three terms of a syl-

[1]These situations are not entirely different from Johnson-Laird's notion of a *m*ental model, which we discuss later. We prefer the term situation so as to not confound the word model, which we take to refer to a computational model.

logism. Situations are sampled from a naive prior. In this setting, it is reasonable to assume that the prior probability, $P(situation)$, is a binomial distribution (i.e. the probability of a coin of a given weight coming up Heads n times in a row, where n is the number of objects). We assume properties are relatively rare of objects; this means the base rate (`br`) of properties is less than 0.50. This principle of rarity comes from the intuition that properties are relative rare [2], and this has been used in probabilistic models previously.

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

To reason over situations, we draw on ideas in formal semantics by assuming that propositions are truth-functional operators. A given proposition (e.g. *A*ll As are Bs is a function which takes in a situation and returns a truth value. From this we get a prior distribution over propositions. By sampling over many situations, we construct a distribution over propositions, shown in Figure 1, first column.

```
(define all (lambda (x) (eval-all x)))
(define some (lambda (x) (eval-some x)))
...
```

The Conditional Semantics model treats the premises of a syllogism as evidence by which to update its belief prior over propositions (which are now potential conclusions). This update rule we take to be probabilistic conditioning. Thus, we get a posterior distribution of propositions, conditioned on premises being true. This has the nice feature of returning a degree of belief in each of the possible conclusions for every syllogism. Using this, we can test the hypothesis that reasoners in syllogistic reasoning tasks are essentially drawing samples from the posterior distribution of the conclusion conditioned on the premises (so-called "posterior matching") (Griffiths & Tenenbaum, 2006).

```
(query
  ...define A,B,C...
  (define conclusion (conclusion-prior))

  conclusion

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

Conditional Pragmatics

One reason to be skeptical of mere *posterior matching* in syllogistic reasoning is that language understanding and language production are central to the syllogistic task, and truth-functional semantics can only go so far. It is natural to think of natural language pragmatics as entering in two places in the model: premise interpretation and conclusion production. We address only production in the current model [3]. To illustrate pragmatic production, we offer as an example the canonical "All/all" syllogism:

All As are Bs
All Bs are Cs

This is in fact one of the easiest syllogisms to which 81% of participants produce the valid*A*ll As are Cs conclusion. The other valid conclusions – *Some As are Cs* and *Some Cs are As* receive only 7% of the endorsements. However, in every possible situation in which *A*ll As are Cs is true, *Some As are Cs* and *Some Cs are As* are also true. That is, they are equally probable given the posterior.

This illustrates one of the issues with the conditional semantics model: it is too literal. The model computes the probability of the conclusion conditioned on the premises being true. For logically valid syllogistic conclusion, this probability is equal to 1. In particular, for any syllogism for which one of the universal quantifiers is valid (i.e. *a*ll and *n*one), the particular quantifier (i.e.*s*ome and *n*ot all, respectively) will also be valid. Human reasoners show a clear preference for conclusions using universal quantifiers (all and none) in these cases.

To address this we draw on recent advances in probabilistic models of pragmatic reasoning. The *rational speech-act theory* addresses the problem of language understanding as that of a speaker trying to convey information about a world or situation that the speaker has observed (Frank & Goodman, 2012). The speaker draws on common-ground of communicative goals to maximize information content of a given utterance. In particular, choosing an utterance in proportion to the likelihood of a listener inferring the situations about which the speaker wishes to convey information. In turn, a listener considers situations in which the given utterance heard is not only true but would be chosen to disambiguate amongst multiple consistent situations by an informatve speaker. This results in the canonical "scalar implicature" wherein the literal meaning of "*some* of the apples are red" is enriched to communicate "some *but not all* of the apples are red".

We use a variant of the nested-conditioning model of scalar implicature (Goodman & Stuhlmüller, 2013) to break the symmetry described above. We treat the reasoner as a sort of pragmatic speaker with a hypothetical listener in the reasoner's mental representation. This listener can be thought of concretely as the person evaluating the participant's responses. In the original scalar implicature model, the listener hears an utterance and tries to reconstruct the situation observed by the speaker. In our model, the listener hears the conclusion and tries to reconstruct the premises with which the reasoner was presented. This distinction turns out to be important as the reasoner does not observe a situation directly but rather constructs situations consistent with the premises.

We speculate that there is another important pragmatic effect which dominates the interpretation of the premises. In

[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.

[3] Preliminary analyses suggest standard Gricean implicatures are not all that's at play during premise comprehension. This is consistent with other empirical findings which set out to test this. We address this matter in more detail later.

what natural setting does a person produce a syllogistic like argument? Why would anyone ever go through the hassle of uttering two sentences which do not convey the totality of what is intended to be conveyed? We suggest this is due to an asymmetry among the three relations (i.e. A-B, B-C, C-A). We posit that the A-B & B-C relations are weighted more strongly than the A-C relationship, possibly owing to common ground of those relationships. In some way, A-C is unexpected, or somewhat unbelieveable. If it weren't, there would be no reason to work through the syllogism. The syllogism was invented as a tool of argument, to convince others. Nobody needs convincing that "Socrates is mortal".

Following up on this intuition, we introduced a small dependency (df) into the binomial prior such that A & C are more likely when B is present and less likely when B is absent. In our example from the beginning, this means that some of your office mates being out of the office for weeks is less likely if they don't have the flu and more likely if they do have the flu.

```
(define B (mem (lambda (x) (flip base_rate))))
(define A
  (mem (lambda (x)
    (flip (if (B x) (+ br df) (- br df))))))
(define C
  (mem (lambda (x)
    (flip (if (B x) (+ br df) (- br df))))))
```

Results

Meta-analysis data

To test our model, we used data from the meta-analysis presented with the Probability Heuristics Model(Chater & Oaksford, 1999). These data were compiled from 5 studies on syllogistic reasoning, completed from 1978-1984. The data included percentage response for conclusions of each of the 4 quantifiers as well as "No Valid Conclusion". We consider the production of the "No Valid Conclusion" an aspect of the algorithmic level of analysis and leave this for future work. Some studies allowed for the conclusion to be ordered in either direction (A-C or C-A) while others restricted it to the classical ordering (C-A). Because of this, we allowed our model to draw conclusions in either order. Following the procedures of the meta-analysis, we collapsed responses across these two orderings to compare it to the data set.

Qualitative results

For each model, we report the total number of syllogisms for which the model's model response is the same as for the meta-analysis. This is a qualitative assessment of the fit.

We first examined the prior to see if it alone accounted for human reasoning patterns. It did not (Figure 1, column 1). Since "not all" is the most likely conclusion to be true, the prior gets only the syllogisms with a "Not all" modal response correct. For the meta-analysis data, this amounts to 29 out of the 64 syllogisms.

When we condition on the premises being true, the model matches people's maximum judgments on 37 of the 64 syllogisms. The 29 syllogisms for which "not all" was the modal response are qualitatively unaffected i.e. the Conditional Semantics model matches these responses. As well, the model matches 8 syllogisms for which "some" and "none" are favored (Figure 1, column 2, see e.g. [2]).

Finally, we introduce pragmatics into the model in part to be able to distinguish among equally "certain" conclusions (e.g. Figure 1, [1]). The model selects not only conclusions likely to be true, but informative conclusions (e.g. Figure 1,[3]). The model now matches 44 out of 64 modal responses.

In addition to capturing many of the modal responses, the model is able to accomodate more than one plausible conclusion. Example [4] in Figure 1 highlights one such example. This is a syllogism with a valid conclusion but one which people find it difficult to draw. The Conditional Semantics 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 relatively more informative than *not all* and so the Conditional Pragmatics enforces this symmetry. Notice how this is different from Example [3], in which *all* is less likely than *some* after conditioning[4]. In this example, *all* gets pragmatically strengthened over and above *some* due *all*'s much higher relatively informativeness.

Still, there are many syllogisms for which reasoning patterns are not captured by the model. A large subset of these are syllogisms which use two negative quantifiers (*not all* or *none*). These problems, in a way, are underconstrained in terms of the situations consistent with the premises. For this reason, the models' predictions do not differ appreciably from the prior (1 [5]).

Model fit

To assess our models' quantitative fits we compute compute correlations across all 256 data points (64 syllogisms x 4 conclusions).

The prior's predictions are the same for all syllogisms and the overall fit is also accordingly poor (r = .36) (Figure 2, row 2, column 1). After conditioning on the truth of the premises, the model is able to make graded responses for each of the syllogisms. These responses are a reflection of the types of situations consistent with the premises. The overall correlation is appreciably higher (r=0.64 (Figure 2, row 2, column 2)). However, among valid conclusions (squares in Figure 2) the fit is terrible (r=0.05). This is a direct consequence of the model's literalness. Among valid conclusions, the model has no preference because each conclusion is true in every possible situation.

This symmetry is broken by using pragmatic reasoning (Figure 2, row 2, column 3). The overall correlation again increases (r = 0.77). The difference is especially striking for the valid syllogisms (r = 0.85). We compared these results to those found when using a uniform prior (2, row 1)) and found the overall fit much better with the binomial, which invokes the principle of rarity.

[4]This may seem surprising as the obvious response is *all*. This is the case because there are two *some* responses which are valid and only one *all* response

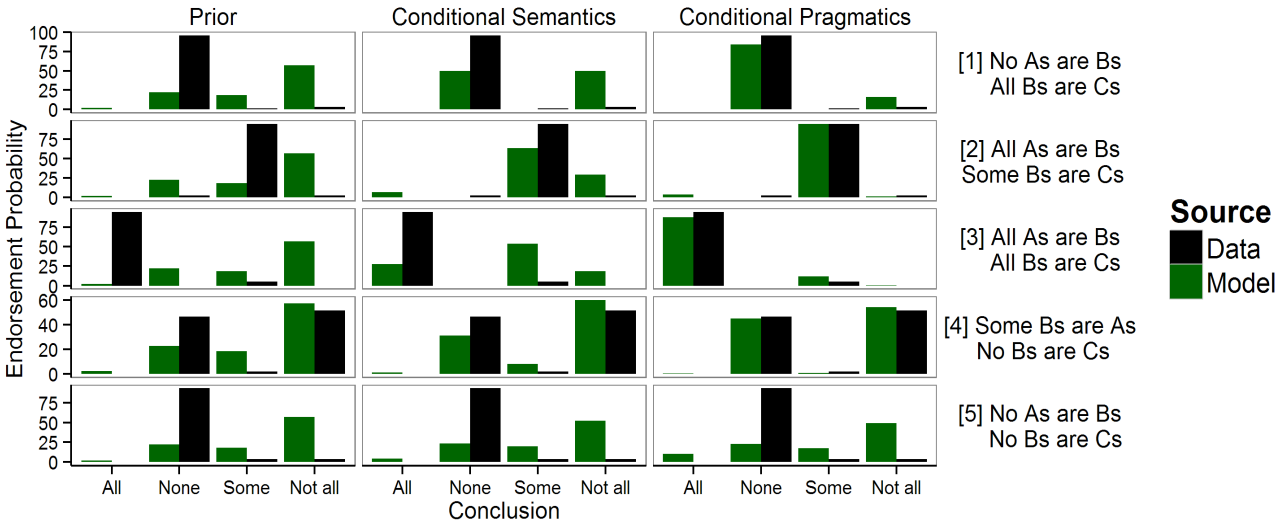

Figure 1: Five example syllogisms. [1] Conditional Semantics has no preference among equally valid conclusions; the symmetry is broken by the pragmatics model which uses includes informativity in its predictions. [2] Semantics alone captures the modal response and pragmatics enriches the quantitative fit. [3] The modal response is not as likely in the Semantics model but is wildly more informative. [4] The models are able to capture multiple preferred conclusions. [5] Models do poorly in matching subjects' responses in an invalid syllogism.

Dependency

The inspiration for this model comes from the idea that syllogistic reasoning can not be disentangled from language understanding. In particular, natural language semantics is insufficient to explain the variability in such a data set. We argue that understanding conversational pragmatics is necessary to understanding how people reason with syllogistic arguments. We have found that the meanings of quantifiers are important insofar as *some* might imply *not all* and so people prefer to conclude *all* when it is warranted.

Using the correlated prior, the Conditional Semantics model gets 45 out of 64 modal responses. The overall fit is also improved, r = 0.75 (Figure 2, row 3). The Conditional Pragmatics model does better as well, predicted the modal response for 52 syllogisms. Additionally, the quantitative fit is high (r = 0.85). Among valid syllogisms, it is correspondingly higher as well (r = 0.88).

Most and few

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 question of the meaning of generalized quantifiers like "most" and "few" is a topic of great interest to the field of formal semantics. Often, "most" and "few" are modeled by a thresholded step 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. We realize this is a gross oversimplification of the meanings of these words. We maintain the set size parameter of 6. However, we feel the usage of the words "most" and "few" might bias people to represent sets of larger size.

We compare our model predictions to two studies carried out by Chater & Oaksford on syllogisms using generalized quantifiers. In these studies, participants were given syllogisms with these generalized quantifiers e.g. *M*ost artists are beekeepers. Few chemists are beekeepers.. Participants were told to indicate which, if any, of the four conclusions followed from the premises and were allowed to select multiple options.

The set of all possible syllogisms with 6 quantifiers contains 144 questions. The authors divided these into two experiments to avoid subject fatigue. Experiment One consisted of the *all*,*not all*,*most*, and *few* quantifiers. Experiment Two used *most*,*few*,*some*, and *none*.

We find good correspondence between the experimental data and the model, even without doing any parameter fitting. The fit is better for the experiment using *all*,*most*,*few*, and *not all* (r = 0.80) than for the experiment using *most*,*few*,*some*, and *none* (r = 0.68). The same was true for the Probability Heuristics Model (r = 0.94 vs r = 0.63). Overall, the proportion of *no valid conclusion* responses, which we do not model, was much higher in Experiment 2 than in Experiment 1. Thus, this data set may well contain more noise than others.

Relationship to previous theories

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

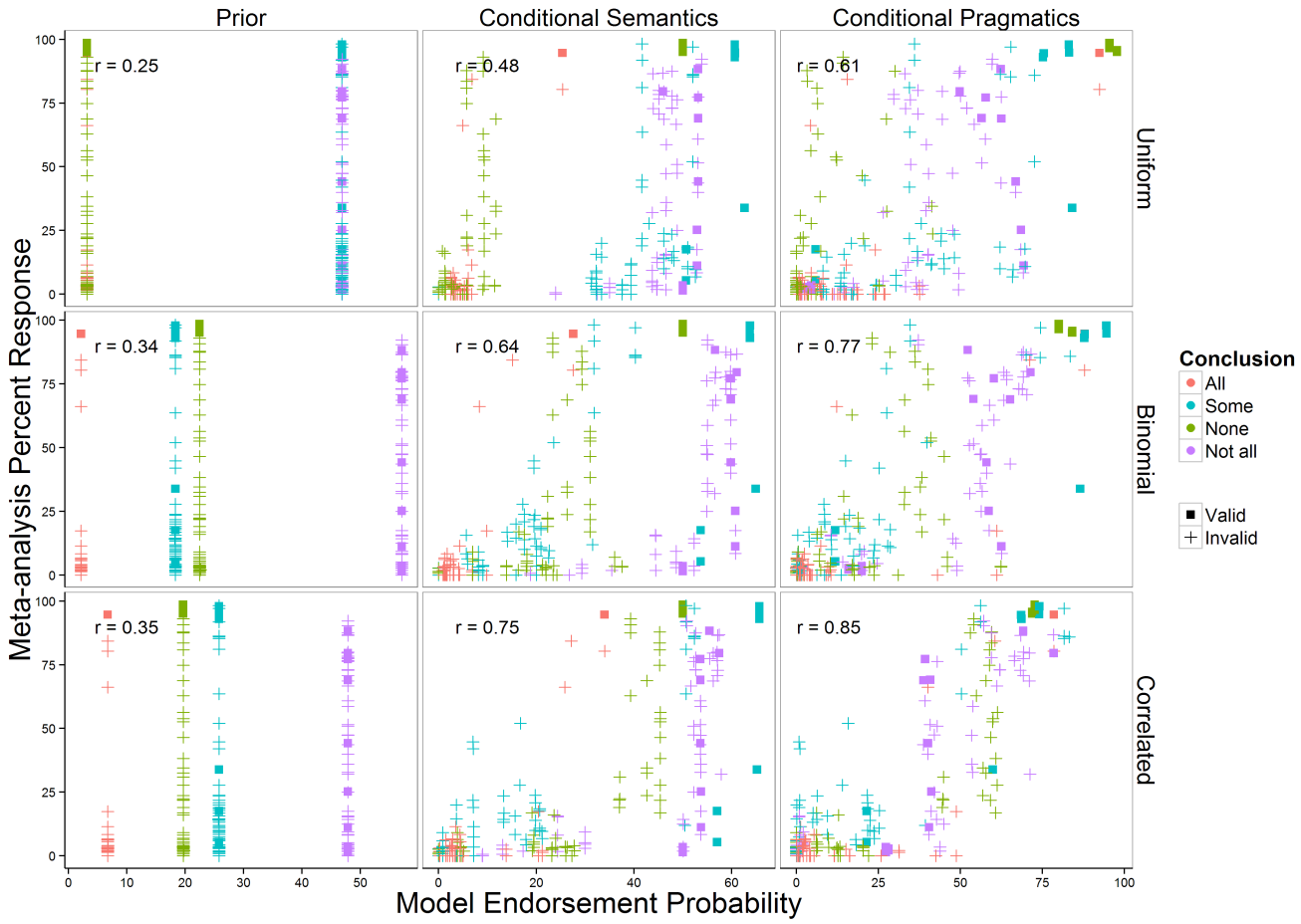

Figure 2: Human subject percentage endorsement vs. model fits for 3 different priors. Columns (from L to R): predictions based only on the prior, Conditional Semantics models, and Conditional Pragmatics models. Rows (from top to bottom): uniform prior, binomial prior, correlated prior (see text).

based on heuristics (Khemlani & Johnson-Laird, 2012). We see the space slightly differently. In one dimension, theories may be based on applying rules – be they heuristic or logical – or they are based on constructing concrete representations or models. In another dimension, theories may be considered fundamentally probabilistic or deterministic.

We now review 2 theories which we take to exemplify two of the four quadrants of this theoretical space.

Mental Models

The Mental Models Theory (MMT) describes a psychological process by which people reason by constructing *iconic* mental representations or models, which represent the terms of a proposition as a collection of individuals. In syllogistic reasoning, a model is constructed for each premise, and premise models are consolidated so that the conclusion may be "read off" the joint model. For example, a model for the premise *All artists are bakers* could be represented as the following situation.

artist	baker
artist	baker
	baker

This shows 2 individuals who are both artists and bakers, and one individual who is a baker but not an artist. Thus, each row is a representation of the properties of an individual. The authors emphasize the need to search for counter-examples to check for logical validity. Errors arise in this search process. Another premise might read: *Some bakers have the flu*. This would like like:

baker	flu
baker	{flu}
	{flu}

The curly brackets reflect our uncertainty about where, if anywhere, to put additional individuals with the flu. The conclusion can be achieved by consolidating these models and reasoning over the joint model.

The MMT captures the intuition that people are able to reason about sets of things explicitly and with respect to context. The a priori believability of propositions has been shown

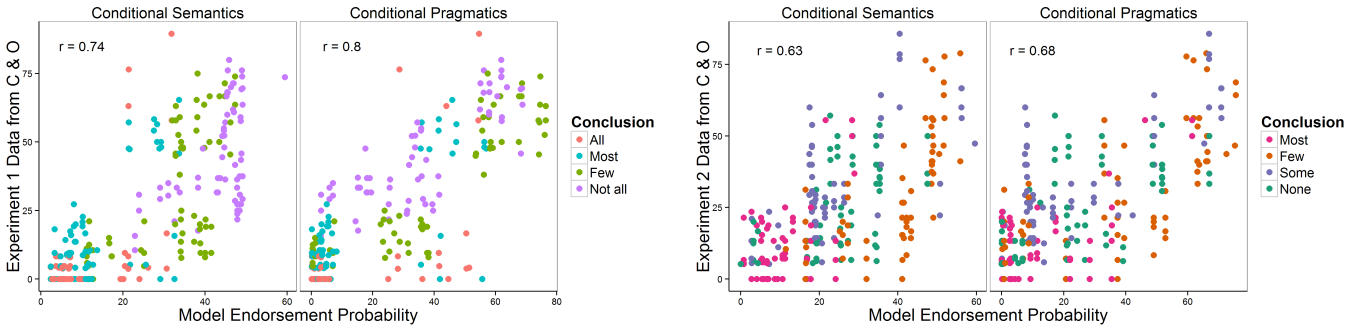

Figure 3: Human subject percentage endorsement vs. model fits for 2 experiments using generalized quantifiers. Experiment 1 (left) used the quantifiers {*all*, *most*, *few*, *not all*}. Experiment 2 (right) used the quantifiers {*most*, *few*, *some*, *not all*}.

to have an important effect on reasoning (Oakhill, Johnson-Laird, & Garnham, 1989). Mental models are flexible in that content can motivate one to carry on the search process longer. The search process allows for individual differences insofar as some individuals may test many models and some may test just one. At the same time, the theory is not well defined insofar as it does not specify how various models come into existence, only that various models can come into existence.

Mental models are very similar to the *situations* described in the Conditional Semantics model. A critical difference is that in Conditional Semantics, situations are construed by sampling. Thus, the model can make quantitative predictions about reasoning patterns with no further assumptions.

Probability Heuristics

We consider heuristic accounts acounts analagous to formal rule accounts in that people are reasoning at the level of propositions, which may be determined by probabilities. This the case with Chater and Oaksford's Probability Heuristic Model (PHM). Like our approach, the PHM is inspired by the notion that people are not fundamentally deductive reasoners, but instead are trying to gage degrees of plausibility for the conclusion. This amounts to computing the probability of a particular conclusion being true given that the premises are true. To accomplish this, the PHM relies on a number of *generation* and *test* heuristics which produce and quantify confidence in conclusions, and which they claim are justified by their computational level theory. The computational level theory includes a notion of informativeness, on which all their heuristics rely. We do not believe their heuristics are necessary for deriving a probabilistic model of reasoning, as we discuss below. Further, as is the case for theories based on formal rules, the very nature of their heuristics suggests reasoners are engaging with the syllogisms at a propositional level and not at the level of concrete representations. We also do not believe this to be the case.

By examining the prior distribution used in the Conditional Semantics model, we can see that all conclusions are not equally likely. In particular, we can consider the reciprocal function of the prior as an ordering of informativity. The Probability Heuristics Model also uses this ordering, though they derive it from propositions, assuming categories are represented as hyperspheres in a high-dimensional concept space (Chater & Oaksford, 1999). In the Conditional Semantics model, this ordering naturally arises from the process of sampling situations.

Discussion

The partitioning described above places the Conditional Semantics and Pragmatics models in the unexplored quadrant of the two-dimensional theoretical space described: we consider reasoning over concrete situations and situations to be constructed probabilistically by sampling.

We have presented a formal model of syllogistic reasoning based on the *rational speech-act framework*. We have shown that in this model, reasoners construct mental situations by sampling, and reasoning over these situations, much like has been described in the Mental Models literature. Unlike Mental Models, however, our model is inherently probabilistic and thus assumes reasoners are in some way gaging degrees of plausibility in syllogistic reasoning tasks. Further, the Conditional Semantics model is fundamentally quantitative, and thus can be considered an elaboration of the Mental Models theory.

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. As well, this model is flexible enough to capture some of the variability in reasoning data using generalized quantifiers. We tested the model assuming a naive threshold of 0.50, such that *most* is true if more than half is satisfied. We think this is an oversimplification and will examine more context-sensitive formulations in future work.

In this framework, a syllogism is read as an argument given as a part of discourse between people. Indeed, this is how syllogisms were used in the time of Aristotle and in the long tradition of scholastic philosophers. Fundamentally, syllogisms are a tool used to convince others. The results of the pragmatic reasoning model shed light on the idea that human rea-

soning behavior in the syllogistic task is as much reason as it is human. Gaging degrees of plausibility alone is not sufficient. A listener needs to be posited at the end of the line so that a conclusion makes sense; so that a conclusion is convincing!

If one accepts that the purpose of a syllogism is to convince, a natural question arises. Why not just assert "Some of my colleagues won't be at work for weeks" from the get-go? Why go through the trouble of laying out premises and having a person draw the conclusion? Arguments are used to persuade, and not all assertions are equally believable a priori. The reason premises are presented in this way is that there is no reason for me to believe some of my colleagues will be out of work for weeks. That almost never happens, except around Christmas break. And so the conclusion is not obvious, and the premises are an alternative route to persuasion. If we accept the premises, we are left with no choice but to draw the conclusion.

We set out to explore this intuition by seeing if the middle term had special significance. The rationale is that if someone is to go through the trouble of mentioning the middle term, it must allow them to assert something that they wouldn't be able to assert otherwise. In other words, we take the a priori probability of A & C to be relatively low, but become higher if B is observed. A & C become correlated via B. This is akin to saying that the end terms have a special relationship via the middle term. It is no coincidence that the same middle term appears in both premises and it is no coincidence the premises appear at all.

We modified the naive binomial prior to induce a slight correlation between A & C via B. Overall, the model matches a few more modal responses and provides a better quantitative fit to the data.

We do not take the probability of co-occurrence to be a priori unlikely in the case of syllogistic reasoning tasks. Many of the experimental materials in the meta-analysis data were well controlled for semantic content. There is no reason to believe *All artists are chemists* is a priori unlikely. Rather, we see this as arising from conversational pragmatics. It is a future direction of this work to develop a formal model of this phenomenon.

‘

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., & Stuhlmüller, A. (2013, January). Knowledge and implicature: modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–84. doi: 10.1111/tops.12007

Griffiths, T. L., & Tenenbaum, J. B. (2006, October). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), 767–73. doi: 10.1111/j.1467-9280.2006.01780.x

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. doi: 10.1037/a0026841

Oakhill, J., Johnson-Laird, P., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition*, *31*, 117–140.

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