Why do you ask? Good questions provoke informative answers.

Robert X.D. Hawkins, Noah D. Goodman
Stanford University

What makes a question useful? What makes an answer appropriate? In this paper, we formulate a family of probabilistic models of question and answer behavior that differ in the amount of pragmatic reflection attributed to speakers. We compare these models based on three different pieces of evidence: First, we explore three classic effects in psycholinguistics that show an answerer's level of informativeness varies with the inferred questioner goal. Second, we jointly test the questioner and answerer components of our model using a simple question-answer communication game. Third, we use a real-time, multi-player version of this game with a wider range of conditions, which allows us to distinguish among the questioner models. We find that sophisticated pragmatic reasoning is needed to account for some critical aspects of the data. People can use questions to provide cues to the answerer about their interest, and can select answers that are informative about inferred interests. The best model accurately fits our behavioral data, providing a promising direction for understanding the dynamics of human question-answer dialog.

Keywords: pragmatics, computational modeling

Introduction

ndg: I moved this to a new file for post-fyp edits (on our way to a journal paper). We should probably move it to a new dir and change the name appropriately....

Q:"Are you gonna eat that apple?"
A:"Oh, go ahead!"

In this exchange, Q wants to know whether she can eat A's apple. Instead of directly asking for the apple, Q strategically chooses a question that differs from her true interest – avoiding an impolite question – yet still manages to send a signal to A about her intentions. A, in turn, reasons beyond the overt question and provides an answer that addresses Q's true interests. This subtle interplay highlights the prevalence of social reasoning tasks in everyday dialogue and raises two specific questions for formal models of language: What makes a question useful? And what makes an answer appropriate?

ndg: I wonder if we should include more intro examples for this longer version? I like this one, but the politeness aspect is slightly misleading....

A number of psycholinguistic studies have provided evidence that answerers are both sensitive to a questioner's goals and attempt to be informa-

This report is based in part on work presented at the 37th Conference of the Cognitive Science Society. The first author is supported by a NSF Graduate Research Fellowship and a Stanford Graduate Fellowship. Correspondence concerning this article should be addressed to Robert X.D. Hawkins, e-mail: rxdh@stanford.edu

tive with respect to those goals. For instance, in Clark's (?) classic study, researchers called liquor merchants and opened the conversation with one of two sentences to set context: "I want to buy some bourbon" (the *uninformative* condition) or "I've got $5 to spend" (the *five dollar* condition). They then asked, "Does a fifth of Jim Beam cost more than $5?" Merchants gave a literal yes/no answer significantly more often in the latter condition than the former, where an exact price was more common. When provided with the five dollar context, the merchant inferred that the questioner's goal was literally to find out whether or not they could afford the whiskey, hence a simple 'yes' sufficed. In the uninformative context, however, the merchant inferred that the questioner's goal was just to buy whiskey, so the exact price was the most relevant response (?, ?).

Context and questioner goals have also been implicated in accounts of answers to identification questions like "who is X?" (?, ?), questions like "Do you have the time?" that permit answers with different degrees of approximation to the true time (?, ?, ?), wh-questions like "Who passed the examination?" where answers can be understood to mean either an *exhaustive* or *selective* list of relevant entities, and to questions like "where are you?" that permit answers at many levels of abstraction (?, ?). This suggests that social inference is a critical aspect of question-asking and answering.

ndg: expand the preceding paragraph to unpack the previous literature on questions a bit more.

Recent work on Rational Speech Act (RSA) models (?, ?, ?) has mathematically formalized pragmatic language understanding as a form of recursive Bayesian inference, where listeners reason about speakers who choose utterances that maximize information gained by an imagined listener. In this paper we extend the RSA framework to address simple question-answer dialogs. The immediate challenge in doing so is that the speaker utility in RSA is based on direct information provided by an utterance—since questions don't provide direct information[1], we must say what utility they do have.

We suggest, following ? (?), that the value of a question is the extent to which it can be expected to elicit useful information later in the dialogue. More specifically, for the questioner, the value of a question is the expected information gained about her interests, given the set of likely answers it may provoke. This diverges from regular RSA in that the value of a question depends on information gained by the speaker (rather than listener), and that this information comes later in the (very short) conversation.

To fully specify this questioner we need a model of the answerer, which can serve as both the model assumed by a questioner, and as a model of answer behavior itself. We explore three, increasingly sophisticated, answerer models. The simplest answerer provides a literal answer to the question (without attempting to be informative); the explicit answerer attempts to be informative with respect to the surface form of the question asked (without inferring the questioner's underlying interests); the pragmatic answerer infers the most likely true interests of the questioner, and then informatively addresses those interests. The latter model uses an extension of RSA to reason about the topic of conversation, as proposed by ? (?); it goes beyond previous work by using the explicit question as a (potentially indirect) cue to this topic.

The rest of this paper is structured as follows. First, we lay out the details of our computational questioner and answer models, discussing similarities and differences with other recent proposals. We then present a set of computational experiments demonstrating how this set of models captures four classic answerer-sensitivity effects from

[1]That is, the literal meaning of a question does not seem to be new information about the world, per se. Questions do, of course, end up conveying information about the speaker's knowledge, needs, and so on—it is this conveyed information that we attempt to derive below.

the psycholinguistics literature. Because data on *questioner* behavior is relatively sparse, and these classic results did not sufficiently engage the questioner component of our model, we collected data in a simple communication game paradigm allowing us to manipulate private goals, potential questions, and potential answers. After testing the predictions of the different models on data from a simple single-player version of the task, we scale up this paradigm to a more natural real-time, multiplayer experiment using a wider variety of goal sets, question sets, and answer sets. We dedicate particular attention in this task to one critical condition in which the three questioner models make different predictions, allowing us to distinguish between them. We close with a brief discussion of how our approach grounds question-asking and answering in social cognition, noting the potential scalability of our framework to problems in natural language understanding and active learning.

A Rational Speech Act model of question and answer behavior

The Questioner

How should a questioner choose between questions? We start by assuming that the questioner aims to *learn information relevant to a private goal*. In order to choose a question that results in useful information, the questioner reasons about how the answerer would respond, given different possible states of the world; she selects a question that results in an answer that tends to provide goal-relevant information.

ndg: somewhere (maybe not here) we need to make it clear that this is a computational level model – we don't assume that people are doing the recursive reasoning online every time they ask a question....

More formally, suppose there is a set of world states $\mathcal{W}$, a set of possible goals $\mathcal{G}$, a set of possible questions $\mathcal{Q}$, and a set of possible answers $\mathcal{A}$. These sets are taken to be in common ground between the questioner and the answerer. An informational goal $g \in \mathcal{G}$ is a projection function that maps a world state to a particular feature or set of features that the questioner cares about; this is similar to the notion of a question-under-discussion (?, ?). We will use the notation $P_g(w)$ to indicate the probability $\hat{P}(g(w))$ of the g-relevant aspect of w under the projected distribution $\hat{P}(v) = \int_{\mathcal{W}} \delta_{v=g(w)} P(w) dw$.

ndg: i think we should revise this notation... e.g. if $P(w)$ is a distribution on $\mathcal{W}$ then write $\widehat{P^g}(w)$ for the probability of $g(w)$ under the projected distribution: $\hat{P}(v) = \int_{\mathcal{W}} \delta_{v=g(w)} P(w) dw$ (which is a distribution over the image $g(\mathcal{W})$).
i also think we should add subscripts of the different model components (Q_0, etc). should make a figure illustrating these components.... we need to unpack and explain the math in this section more (since we have room in the long paper version).

The **questioner** takes a goal $g \in \mathcal{G}$ as input and returns a distribution over questions $q \in \mathcal{Q}$:

$$P(q|g) \propto e^{\mathbb{E}_{P(w^*)}\left[D_{KL}(P_g(w|q,w^*) \,\|\, P_g(w))\right] - C(q)}$$

ndg: hmm... i think this isn't quite right: as defined P_g isn't a distribution (on W). i think what we want is the KL of the g-projections of each distribution on W. this then raises a question about whether the KLs for different goals are comparable in the right way – e.g. smaller g-images have less capacity for information gain – is this right?

It trades off the cost of asking a question, $C(q)$, and expected information gain. The cost likely depends on question length, among other factors. Information gain is measured as the Kullback-Leibler divergence between the (g-relevant) prior distribution, $P_g(w)$, and the posterior distribution one would expect after asking a question q whose answer reflected true world state w^*:

$$P_g(w|q, w^*) = \sum_{a \in \mathcal{A}} P_g(w|q, a) P(a|q, w^*)$$

This distribution has two components: First, it depends on $P(a|q, w^*)$, a model of the answerer which we will explore shortly. Second, it depends on (the goal projection of)

ndg: I don't think we want the goal projection here. we want the distribution over worlds, in order to get a posterior distribution on worlds. we will then project along g in the KL operator above.... (note: need to check against implementation.)

$P(w|q, a)$, an 'interpreter' that specifies the likelihood assigned to different worlds given question and answer pairs.

To define the interpreter function, which all agents use to compute the literal interpretation of a question-answer pair, we must assign questions a semantic meaning. We assume that a question is an informational goal that projects from worlds to the answer set $\mathcal{A}$. This is equivalent to the more common partition semantics of ? (?), as can be seen by considering the pre-image of such a projection; an answer picks out an element of the partition via $q^{-1}(a)$. The **interpreter** constrains the prior on worlds to the subset of its support that is consistent with the semantics of a question-answer pair[2]:

$$P(w|q, a) \propto P(w)\delta_{q(w)=a}$$

The Answerer

How should an answerer choose between answers to a question? What should a questioner assume about the answerer when choosing a question? We next describe three different answerer models; the questioner could assume any one of them, leading to three corresponding versions of the questioner model. All answerers take a question $q \in \mathcal{Q}$ and a true world state $w^* \in \mathcal{W}$ as input and return a distribution over answers $a \in \mathcal{A}$. The **literal answerer** simply chooses answers by trading off prior answer probability and how well a question-answer pair conveys the true state of the world to an interpreter:

$$P(a|q, w^*) \propto P(a)P(w^*|q, a)$$

For a fixed question, this is equivalent to the speaker in previous RSA models. The question enters only in specifying the literal meaning of an answer (e.g. "yes" and "no" pick out different worlds in response to different questions). The **explicit answerer** additionally evaluates answers with respect to how well they address the explicit question q:

$$P(a|q, w^*) \propto P(a)P_q(w^*|q, a)$$

This uses the question utterance itself as a projection operator on the world state.

The **pragmatic answerer** also evaluates answers with respect to how well they address the questioner's goal, but doesn't take the question's explicit meaning at face value. Instead, the pragmatic answerer reasons about which underlying goals g are likely given that a question q was asked, and chooses answers that are good on average:

$$P(a|q, w^*) \propto p(a) \sum_{g \in \mathcal{G}} P(g|q)P_g(w^*|q, a)$$

Reasoning backwards from questions to goals is a simple Bayesian inversion of the (explicit) questioner using a prior on goals:

$$P(g|q) \propto P(q|g)P(g)$$

ndg: need to say here that we could depend on a sophisticated questioner model here, but we use the simplest(?) one for simplicity....

For all of the questioner and answerer models, we can vary how strongly optimizing they are—that is, to what extent they are sampling from the distributions defined above, and to what extent they deterministically choose the most likely element. For any such distribution over utterances, we introduce an optimality parameter α and transform it by $P'(x) \propto P(x)^\alpha$.

[2] In a complete language understanding model we would also have a semantic evaluation function that maps an answer utterance to its value in $\mathcal{A}$. For clarity we assume this is a trivial mapping and suppress it.

This concludes our specification of the model space, giving a set of three answerers and three corresponding questioners that reason about them. We have implemented these models in WebPPL, a probabilistic programming language (?, ?), and runnable code for all reported simulations is available online at `http://hawkrobe.github.io/Q_and_A/`. The model predictions shown throughout the rest of the paper are computed using this implementation.

Related models

ndg: i wonder if we should move this to a section on background just after the intro, that has a subsection for experimental work and one (this) for formal models? we can use that to set up our model.... need a clearer explanation of why van rooij isn't enough / why it needs to be merged into the RSA framework.

The above probabilistic model of question and answer behavior bears some resemblance to recent decision theoretic (?, ?) and game theoretic (?, ?) models of pragmatic reasoning in language use. In particular, all these approaches emphasize *inference* about a partner's underlying mental state. In this respect, these theories contrast with the interactive alignment model (?, ?) and competing dynamical systems models of dialogue (e.g. ?, ?), where coordination occurs through a low-level process of priming and adjusting to a partner's syntactic, lexical, and phonological choices.

While there is a growing literature on dialogue models in general, far less attention has been devoted to the specifics of question and answer dynamics *per se*. The primary theoretical work in this area derives from formal linguistic theories, which focused on the notion of informativeness. In Groenendijk and Stokhof's(?) foundational work on question and answer semantics, asking a question induces a partition over the space of possible worlds, where each cell of the partition corresponds to a possible answer. An answer, then, consists of eliminating cells in this partition, and the most useful answers are those that eliminate all relevant alternatives to the true world. However, as van Rooy (?) and others (?, ?) have pointed out, this predicts that wh-questions like "Where can I buy an Italian newspaper?" can only be fully resolved by exhaustively mentioning whether or not such a newspaper can be bought at each possible location. Clearly, this is not the case: a single nearby location would suffice. These theories also cannot account for contextual variation in what counts as a useful answer.

More recent theories have tried to fix these problems by introducing some consideration of the questioner's goals. ? (?), for instance, formalizes these goals as a decision problem faced by the questioner. A useful answer under this decision theoretic account is one that maximizes the expected value of the questioner's decision problem. A useful question is one that induces a sufficiently fine-grained partition, optimally distinguishing the worlds relevant to the decision problem. While this framework elegantly accounts for the context-dependence and relevance-maximization of question and answer behavior, it assumes that the questioner's decision problem is known a priori by the answerer. If this were the case, the act of asking questions would seem irrelevant: why wouldn't the answerer directly tell the questioner which action to take?

ndg: che

Our models are an effort to expand on this core idea in a probabilistic framework, which also provides a mechanism for the answerer to *infer* the 'decision problem.' The RSA framework has already been applied to account for diverse pragmatic phenomena including scalar implicature (?, ?), interpretation of context-sensitive adjectives like "tall" or "cheap" (?, ?), non-literal language use like hyperbole (?, ?) and irony (?, ?), and acquisition (?, ?). Our model situates question-answer behavior within this unifying view of language understanding as social inference, allowing us to formulate and evaluate competing hypotheses about what that inference entails. We now proceed to show how our model addresses four classic

examples of question and answer pragmatics from the psycholinguistics literature.

Four case studies

Clark (1979), Experiment 4

This example is redundant with the next one – both simply make the QUD prior dependent on a context statement. Maybe we should leave this one out, since we're already using the other Clark experiment?

First, we show that our model can provide different—sometimes over- or under-informative—answers to the same explicit question, depending on context. For this illustration, we model the whiskey-pricing study presented in the Introduction (?, ?). Recall that liquor merchants were more likely to give over-informative answers (specifying exact price) to the question "Does a fifth of Jim Beam cost more than $5?" in the uninformative context ("I want to buy some bourbon") than in the five dollar context ("I've got $5 to spend").

Our world state is simply the whiskey's price ($1, $2, ..., $10). There are two possible goals: learning the price of whiskey and learning whether the price is greater than $5. The set of answers includes exact prices as well as "yes" and "no", with lower cost for "yes" and "no" than the price statements. We model the context sentence as affecting the questioner's goal prior. When the context is "I'd like to buy some whiskey," we assume that the the two goals are equally likely. When it is "I only have $5 to spend," we assume that it is 9:1 in favor of learning whether the price is greater than $5.

Results. When the question is "Does Jim Beam cost more than $5?", the correct Boolean answer is the most probable choice (at probability .44 and .49). Critically, there is a context-dependence for answers to this question: when prefaced with "I'd like to buy some whiskey.", the correct exact price answer is favored more strongly (at probability .18) than when the context is "I only have $5 to spend." (probability .11). By contrast, the literal and explicit answerers (which have no natural way to account for context) do not make differential predictions in the two situations. The literal model predicts that the answerer is equally likely to say the true Boolean answer and the true numerical answer, and the explicit model predicts that the answerer will always give the true Boolean answer, since it is the explicit question being asked. This suggests that our pragmatic *answerer* is consistent with human behavior in psychologically interesting situations, passing a first, qualitative, test.

Groenendijk and Stokhof (1984)

For a slightly more complex example, we consider the classic puzzle of *mention-some* and *mention-all* readings of wh-questions (?, ?, ?). Some questions, like "Who is coming to dinner tonight?" are intended to elicit an exhaustive list of the entities that are answers. For other questions, like "Where can I find a bathroom in this building?", a single answer would be sufficient.

The question "Where can one buy an Italian newspaper?" can intuitively be ambiguous between these meanings depending on who is asking: if it is a tourist, they probably just want to know the nearest place, but if it is a businessperson trying to build a newspaper distribution network in town, they likely want the whole list. The puzzle is determining how the same question can take on different semantics in different contexts: according to our account, this happens via an inference about the questioner's underlying goal.

Our world state is an object consisting of four cafes in town. Each cafe is assigned two properties: its distance from the speaker and whether or not it sells Italian newspapers. There are two possible goals: learning the identity of the *nearest* cafe selling a newspaper and learning the identity of *all* cafes selling a newspaper. The set of answers includes all 16 combinations of different cafes (e.g. "cafe 1 and cafe 3" or "cafe 2, cafe 3, and cafe 4"), as well as the answer "none".

The prior over answer utterances is constructed

as follows: there is a 10% chance of saying "none," to allow for this outcome. Otherwise, the agent selects one of the four cafes and terminates with probability .5. If the agent does not terminate, they pick another cafe from the list and continue until either terminating or running out of possible cafes. This naturally gives longer answers lower probability, reflecting their higher cost of utterance. We model the context sentence as affecting the questioner's goal prior: if they say "I'm new here," there is a 9:1 chance that they are interested in the closest location with a newspaper; if they say "I'm a businessperson...," there is a 9:1 chance that they are interested in all of the newspaper locations.

Note that this formulation of the problem differs slightly from the one given by ? (?), although they easily map onto one another. In the original formulation, the *answer* is fixed and the *meaning* of the answer must be inferred to either be mention-some or mention-all. In our formulation, the meaning of each answer is unique and fixed, but the answerer must choose among a set of different possible answers. The same inferential machinery that our questioner uses to reason about which answer utterance the answerer will give could also be used to reason about which *meaning* the answerer intends by their fixed utterance. The only difference is that the questioner would have uncertainty over a set of possible meanings instead of uncertainty over a set of possible answer utterances. We chose the latter because it more closely resembles the other scenarios we are modeling.

Results. For concreteness, we set the world to be the following :

```
world = {'cafe1' : [3, false],
         'cafe2' : [1, true],
         'cafe3' : [3, true],
         'cafe4' : [3, true]}
```

and enumerated over all model executions for both contexts. We find that the highest probability response given the "I'm new here" context is the single location "cafe2," with probability 0.56. Given the "I'm a businessperson..." context, however, the highest probability response is the conjunction "cafe2 and cafe3 and cafe4," with probability 0.82. Note that cafe 1 was not assigned any probability in either context because it did not sell Italian newspapers (demonstrating that the pragmatic answerer will not lie), and that the nearest cafe was given precedence in the "tourist" context.

The literal and explicit answerers, as in the Clark scenario, do not have the means of making inferences about the questioner's underlying goals. Thus, both models incorrectly predicted that the context would not affect the preferred response. The literal answerer predicted that all combinations of cafes 2, 3, and 4 would be given in proportion to their prior likelihood (with no special preference given to the closest one). The explicit answerer predicted that cafe2 would be preferred in all contexts. Crucially, the only difference between our model of this scenario and the Clark scenario is the set of QUDs, the structure of the world, and the meanings of the answers. The questioner and answerer functions stayed exactly the same.

Gibbs Jr. & Bryant (2008): Experiment 3

It has been shown in previous work that people typically round their answers to the nearest 5 or 10 minute interval when asked 'Do you have the time?", even when they're wearing a digital watch (?, ?). ? (?) replicated this result, and then performed a follow-up study where they preceded their question by the context "I have a meeting at 4:00." They found that the tendency to round times decreased as a function of the time remaining until the stated deadline: when people were asked at 3:40, they would say "It's 3:45," but when people were asked at 3:53, they would make their response precise to the minute. They explain this result by appealing to the questioner's goals: while an approximate time is sufficiently informative with respect to most goals, a questioner who is running late to an appointment may have the goal of judging whether to rush, in which case a precise time is needed.

We take the world state to be the true current

time. Unlike in the previous two scenarios, where there are only two QUDs and two contexts that make each QUD more or less likely, we now let there be a *family of QUDs* and a fixed context stating the time of the appointment. This family of QUDs is parameterized by a threshold time, below which times are rounded to the nearest 5 minute increment and above which times are given exactly. For example, the QUD with the threshold set at 3:50 corresponds to a goal of wanting to learn the true time if it is higher than 3:50 and only caring about the approximate time below 3:50. The prior over thresholds (τ = 3:30, 3:35, ..., 3:55), given a context sentence, weights each threshold by its distance to the appointment time, reflecting the intuition that the questioner will be more interested in the exact time as it approaches the appointment time.

The set of answers is the set of times that could be given (e.g. 3:30, 3:31, ..., 4:00), with uniform probability. Critically, the meaning of an answer is not exact: it introduces a factor in the "interpreter" component of the model, upweighting world states that are close to the answer utterance. In other words, we specify an "approximate" semantics for number references, which is necessary for rounded answers to be produced in the first place.

Results. To test the behavior of this model, we compare two simulations. In both, the context is "I have an appointment at 4:00." In the first, the true world state is 3:34 (the 'appointment far' condition); in the second, the true world state is 3:54 (the 'appointment near' condition). Gibbs Jr. and Bryant (?) found that the answerer is most likely to round to 3:35 in the 'appointment far' condition, but most likely to give the exact time, 3:54, in the 'appointment near' condition. Indeed, in the 'appointment far' condition, the pragmatic questioner model, which infers which threshold is likely given the context and threshold, replies the rounded time '3:35' with probability 0.40 compared to the true time '3:34' with probability 0.27. In the 'appointment near' condition, this pattern reverses: the pragmatic questioner model replies the rounded time '3:55' with probability 0.16 and the exact time '3:54' with probability 0.68 (see Figure **??**).

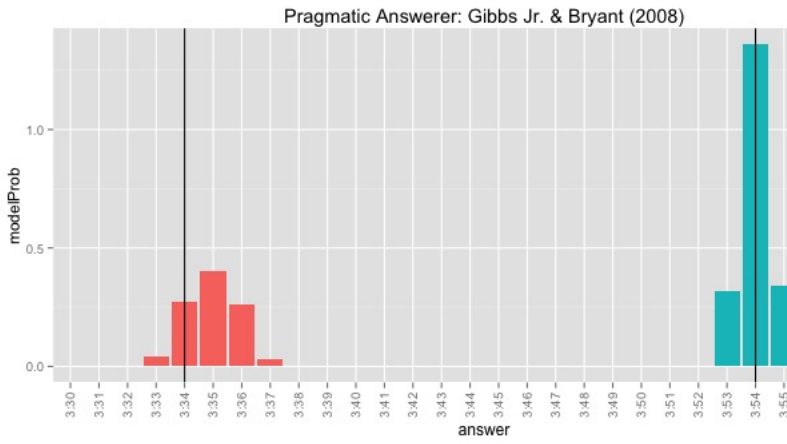

Figure 1. Results for our computational experiments replicating Gibbs Jr. and Bryant (2008). Vertical lines represent the true world state.

Clark (1979): Experiment 5

Our final scenario is a critical test for the questioner component of our model. Because the question space in the previous computational experiments only contained one element, the pragmatic answerers' inferences were entirely based on the context and the QUD prior. One of the most interesting and novel predictions of the RSA model, however, is that the questioner's choice of utterance itself should guide a pragmatic answerer's inferences about likely underlying goals. While there are few experimental results using questioner behavior as the *dependent variable*, making it difficult to test our model's predictions about questioner behavior, there is some work using the question asked as an *independent variable* and testing how it affects answers.

One such study was conducted as a follow-up to the first experiment we modeled in this paper. Instead of calling liquor merchants, ? (?) called restaurants and asked one of four yes/no questions about which *credit cards* the restaurant accepted:

1. “Do you accept Master Charge cards?”

2. “Do you accept American Express cards?”

3. “Do you accept credit cards?”

4. “Do you accept any kinds of credit cards?”

He analyzed the likelihood that the respondent gave a yes/no answer, compared to the likelihood of giving the full (over-informative) list of exactly which cards were accepted. He found that (1) and (2) were nearly always answered with a ‘yes’ or ‘no’, (3) was equally likely to be answered with yes/no and full information, and (4) was nearly always answered with full information.

We formalize this scenario in our model in the following way. The set of possible worlds is given by an object mapping five differents kinds of cards (‘Visa’, ‘MasterCard’, ‘American Express’, ‘Diner’s Club’, and ‘Carte Blanche’) to a Boolean representing whether or not they are accepted. The true world state is known by the answerer but not the questioner. The above four questions form the question space, and the answer space contains ‘yes’, ‘no’, and all possible combinations of cards, including the empty list.

There are four possible goals, which we can understand through four possible questioner situations. First suppose that the questioner only has MasterCard. Thus, a full list would not be helpful – they only want to know whether or not MasterCard is accepted. This is the “MasterCard” goal. Second, suppose that the questioner has both MasterCard and Diner’s Club. Then they want to know whether the restaurant takes *either* card. This is the “Master + Diner’s” goal. Third, suppose that the questioner has MasterCard, Diner’s Club, and American Express. Then they just need to find out whether any one of these cards is accepted. This is the “Master + Diner’s + American” goal. Finally, they might be interested in learning the actual set of names of cards accepted by the restaurant. This is the “names” goal.

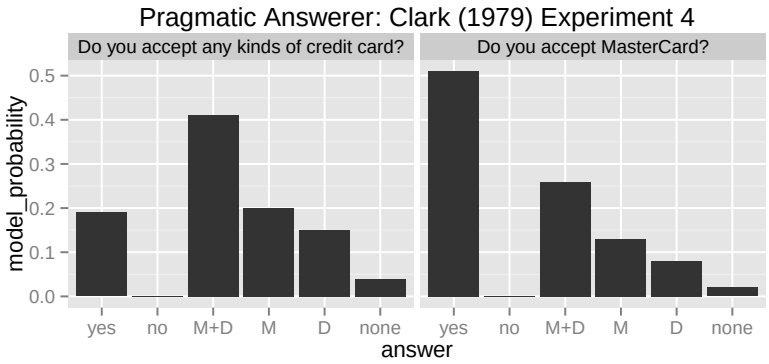

Figure 2. Results for our computational experiments replicating Clark (1979), Experiment 4. M stands for the “MasterCard” answer, D stands for the “Diner’s Club” answer, and $M + D$ stands for the answer mentioning both. Note that we give an over-informative answer to “Do you accept any kinds of credit card?” in the left panel, but a literal ‘yes’/‘no’ answer to “Do you accept MasterCard?” in the right panel.

Results. We tested the behavior of our pragmatic answerer on the following world state:

```
var world = {
      'Visa' : false,
      'MasterCard' : true,
      'AmericanExpress' : false,
      'Diners' : true,
      'CarteBlanche' : false
};
```

We find that when the questioner asks “Do you accept MasterCard?”, the pragmatic answerer is most likely to respond ‘yes,’ but when the questioner asks “Do you accept any kinds of credit card?”, they are most likely to respond with the full list of credit cards accepted (see Fig. **??**).

When we examine the inferred QUD in each case, these results become clearer. For the “MasterCard” question, the pragmatic answerer reasons that the most likely goal is to specifically find out about MasterCard, so “yes” is a sufficient response. For the “any kinds” question, however, the pragmatic answerer reasons that the most likely

goal is to get the full list of names. "Yes" would not adequately address this underlying goal, so the answerer chooses to give the full list instead. Note that these inferences are made purely on the basis of the questioner model's behavior rather than general context as in the previous simulations.

I'd like to extend this example a step further by having the restaurant-owner reason about *which cards the caller owns*, with the assumption that asking about a larger set of cards requires the questioner to pay a higher cost (thus the "any cards" question). Having a possible goal be "learning the full set" seems a bit of a stretch.

Discussion

Remarkably, the same questioner and answerer program was able to reproduce patterns of question-answer behavior in four different scenarios. It captured both explicit and implicit context effects as well as effects where the question itself served as a signal about the relevant underlying goals. Note that all of these studies were focused primarily on answerer behavior, allowing us to show that our model of the questioner is consistently used as an submodule of the answerer. However, because questioner behavior was always manipulated as an independent variable, we could not test the questioner model as a stand-alone predictor of human questioning behavior. This reflects a general neglect of questioner behavior in the psycholinguistics literature, and further experiments are needed to test its predictions.

There remain some deep questions about the behavior of our model in these scenarios, particularly the final one. Recall that Clark found a difference between "Do you accept any kinds of credit cards?" and "Do you accept credit cards?" Our model treats these two questions as equivalent (with the literal semantics returning true if any of the Booleans in the object are true and false otherwise), which propagates all the way up to the pragmatic answerer. It is not obvious where the asymmetry arises. Another issue is that our resulting answer distribution does not appear to be robust to all parameter settings and worlds. The answer prior uses a parameter to determine the likelihood of 'yes' and 'no' responses versus lists of card types, and different settings of this parameter yield very different absolute answer distributions (although they all appear to shift toward fewer 'yes'/'no' answers and more list answers in response to the "any kinds" question). It will be valuable to conduct a more rigorous exploration of the model space, and also test sensitivity to QUD alternatives.

While the questioner component of our model was critical to the pragmatic answerer's behavior in the final simulation, none of these scenarios truly tested the questioner component. Indeed, most empirical work has treated the question or context as an independent variable, with the answer as a dependent variable. Our model, following ? (?), however, suggests that the question itself is important in prompting a relevant answer. In order to test these predictions, we designed a sequence of experiments using a communication game to collect data on both question-asking behavior *and* question-answering behavior.

Exp. 1: Hierarchical questions and answers

Experiment 1 and 2 are the single-player versions, which I think we can take out in the version we submit to a journal – the first multi-player exp (#3) is just a better version of experiment 1, and the second multi-player one (#4) succeeds in distinguishing the questioner models much better than experiment 2, which is quite confusing.

In order to simultaneously test how questioners choose questions when faced with a particular goal and how answerers respond under uncertainty about this goal, we used a guessing-game task played by two players: a questioner and an answerer. In this game, 4 animals (a dalmatian, a poodle, a cat, and a whale) were hidden behind 4 gates. These animals corresponded to differ-

ent levels in a class hierarchy (see Fig. **??**). The questioner received a private goal of finding one of the objects (e.g. 'find the poodle'), and the answerer (but not the questioner) knew the location of each object. Before choosing a gate, the questioner asked the answerer a single question, chosen from a restricted set of options, and the answerer responded by revealing the object behind a single gate. This restriction was motivated by one of the key features of our opening example: when the most direct question ("can I eat your food?") is suppressed due to politeness, utterance length, complexity, or some other intervening factor, questioners must rely instead on an indirect question.

This set of restricted options was critical to distinguishing between the pragmatic and explicit variants of our model. If all questions were equally available, both our 'explicit' and 'pragmatic' questioner models would prefer the most direct one. To see how they make different predictions in the presence of restrictions, suppose 'poodle?' was not available the questioner. If the questioner asked about a 'dog?', the poodle and dalmatian would be considered equally good options by an explicit answerer because they are both dogs. However, the pragmatic answerer could reason that if the questioner was truly interested in the location of the dalmatian, he would have *asked* about the dalmatian. Because he didn't, he must be interested in the other valid response that he lacks a direct question for: the poodle.

Participants. We recruited 125 participants from Amazon's Mechanical Turk to participate in this task. Eleven participants were excluded due to self-reported confusion about the task instructions or due to being non-native English speakers.

Stimuli & Procedure. In terms of our model specification, the world space $\mathcal{W}$ was the set of 4! = 24 possible assignments of four objects to four gates. The goal space $\mathcal{G}$ was the set of four objects that the questioner could be trying to find (the leaves of the tree in Fig. **??**). The answer space $\mathcal{A}$ was the set of four gates that the answerer could reveal. The restricted question space $\mathcal{Q}$ contained

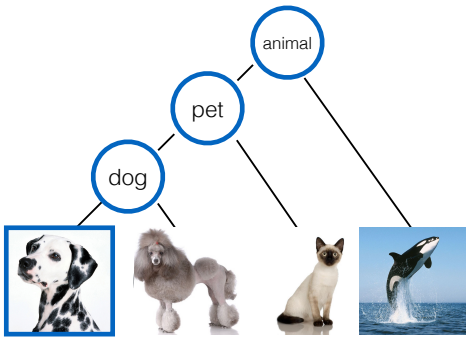

Figure 3. Stimulus hierarchy used in Exp. 1. The goal space and answer space contained the four leaves. The question space, however, was restricted to the highlighted nodes, proceeding up the hierarchy, allowing for indirect questions.

the set of highlighted nodes in the hierarchy: 'dalmatian?', 'dog?', 'pet?', and 'animal?'.

Each participant provided responses for four trials in the role of the questioner (corresponding to the four goals), and four trials in the role of the answerer (corresponding to the four possible questions). In the questioner block, players were presented with a private goal from $\mathcal{G}$, like "find the poodle!" and were prompted to select a question from a drop-down menu containing elements of $\mathcal{Q}$ that would best help them find it. In the answerer block, players were shown which items were behind which gates and were told that the other player had asked a particular question from $\mathcal{Q}$. They were prompted to select a gate from a drop-down menu that would be most helpful for the questioner, keeping in mind his or her constraints. (To minimize learning effects, questioners did not receive answers and neither role saw the outcome of the game.) In order to collect responses for all elements of $\mathcal{G}$ and $\mathcal{Q}$, the order of the questioner and answerer blocks was randomly assigned for each participant, and the order of stimuli within these blocks was also randomized

Results. Results for the questioner role are shown alongside model predictions in Fig. **??** (left). We find that questioners systematically prefer to ask different questions given differ-

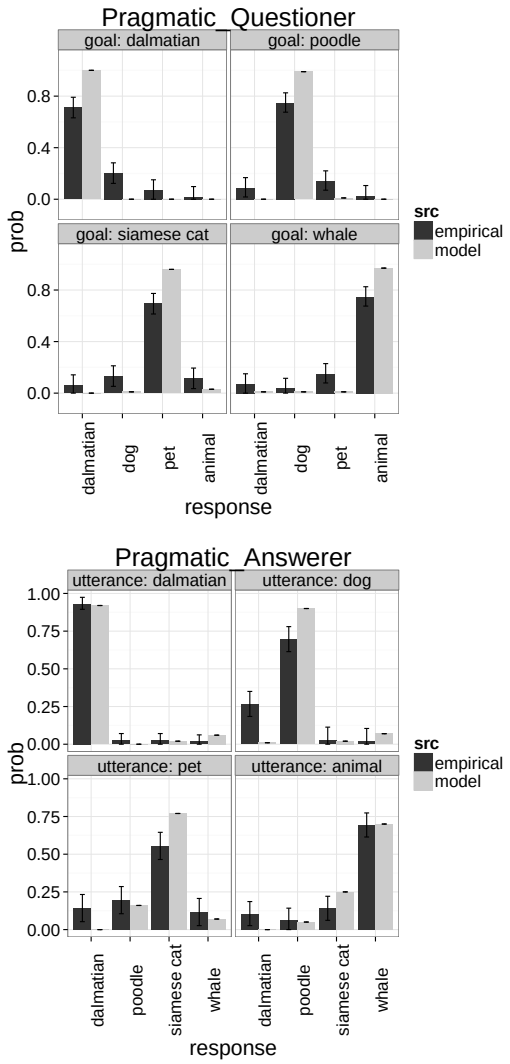

Figure 4. Exp. 1 results, compared with the predictions of the best-performing model for questioner (left) and answerer (right). The explicit and pragmatic questioner models do not make different predictions in this task, but the pragmatic answerer better accounts for the qualitative patterns in the response data than the explicit answerer.

ent goals, even as those questions become more indirect. χ^2 tests over each of the four response distributions show a significant divergence from uniform. Questioners preferentially ask about the 'dalmatian' given the dalmatian goal, $\chi^2(3) = 137, p < .001$, about the 'dog' given the poodle goal, $\chi^2(3) = 152, p < .001$, about the 'pet' given the cat goal, $\chi^2(3) = 120, p < .001$, and about the 'animal' when given the whale goal, $\chi^2(3) = 150, p < .001$.

Results for the answerer role are shown in Fig. **??** (right). Answerers are highly sensitive to the constraints of the questioner, giving information about the dalmatian when asked about a 'dalmatian', $\chi^2(3) = 281, p < .001$, about the poodle when asked about a 'dog', $\chi^2(3) = 137, p < .001$, about the cat when asked about a 'pet', $\chi^2(3) = 57, p < .001$, and about the whale when asked about an 'animal', $\chi^2(3) = 121, p < .001$. Note that, under an explicit interpretation of the question, revealing the dalmatian and the poodle would both be perfectly acceptable answers to a question about a 'dog', but answerers strongly prefer to give the location of the poodle. In the next section, we compare these results to the predictions of our family of models (Fig. **??**).

Model comparison. Each model was run with uniform prior probability over worlds, goals, questions, and answers, and with equal cost for all utterances. For each model, a single optimality parameter, which applied to all agents as described above, was fit to maximize correlation with the data.

We can rule out both the literal answerer and literal questioner. The *literal answerer* yields a uniform distribution over the four answers. This has consequences for the corresponding *literal questioner* model: when this questioner reasons about which question would generate the most helpful answer from the literal answerer, it finds no differences in response probabilities, and therefore has no preference for which question to ask. The predictions of these model, plotted against our empirical results, are shown in the left-hand column of Fig. **??**.

The two remaining questioner models make roughly the same predictions for this task, and we are not able to distinguish them on the basis of these data. We found a model-data correlation of $r = 0.96$ for the explicit questioner and correlation of $r = 0.99$ for the pragmatic questioner. The difference between these correlations is signif-

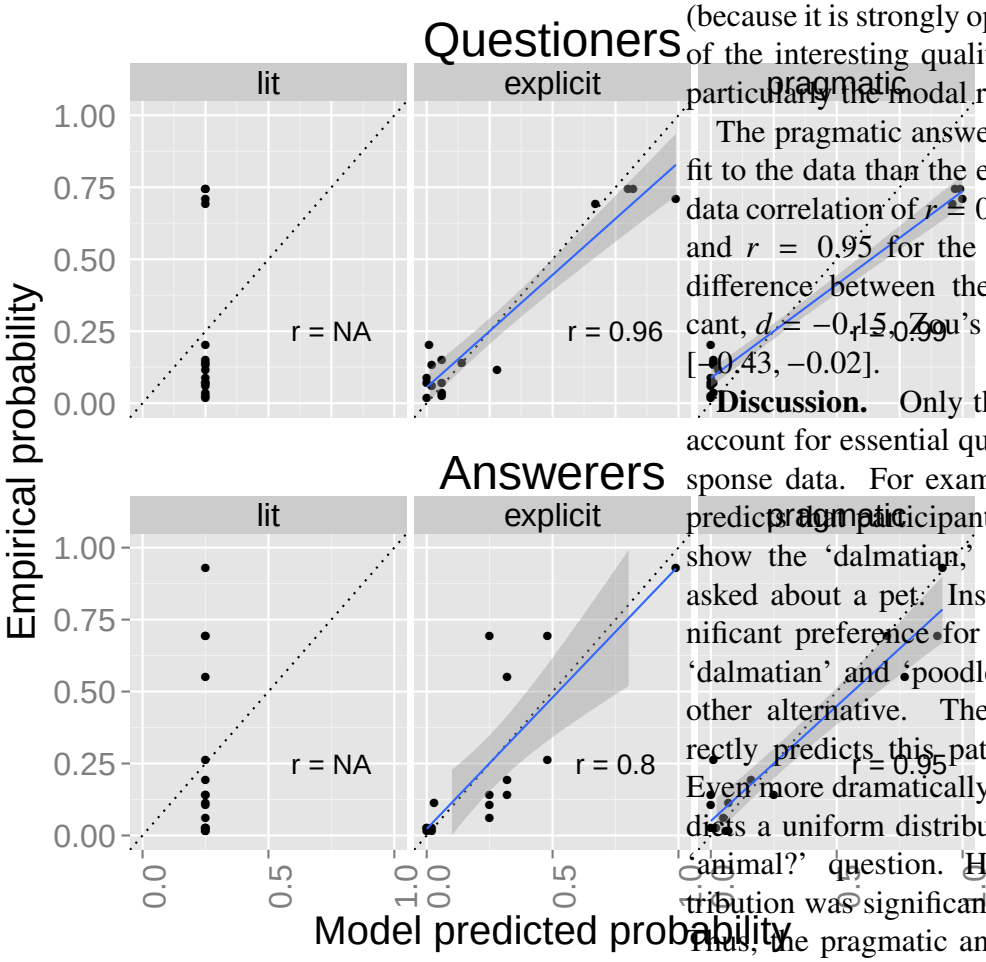

Figure 5. Full space of models, and their correlations with the data from Exp. 1. Questioner models in the first row reason about the answerers directly below them, and the pragmatic answerer reasons about the explicit questioner.

icant, $d = 0.03$, hou's 95% confidence interval = $[-0.097, -0.004]$, taking into account the fact that these correlations are dependent and overlapping on the same empirical data (?, ?). Although the pragmatic model has a slightly better fit, the two models only differ slightly in the magnitude of predictions, not in qualitatively important ways such as the rank ordering of response. The pragmatic questioner model's predictions for each response distribution are shown in Fig. **??** (left). Although the magnitude of its predictions are not in perfect alignment with the magnitude of the empirical data (because it is strongly optimizing), it captures most of the interesting qualitative patterns of the data, particularly the modal responses.

The pragmatic answerer provides a much better fit to the data than the explicit answerer: a model-data correlation of $r = 0.8$ for the explicit answerer and $r = 0.95$ for the pragmatic answerer. The difference between these correlations is significant, $d = -0.15$, hou's 95% confidence interval = $[-0.43, -0.02]$.

Discussion. Only the pragmatic answerer can account for essential qualitative features of the response data. For example, the explicit answerer predicts that participants will be equally likely to show the 'dalmatian,' 'poodle,' and 'cat' when asked about a pet. Instead, the data show a significant preference for revealing the cat, leaving 'dalmatian' and 'poodle' at the same level as the other alternative. The pragmatic answerer correctly predicts this pattern (see Fig. **??** (right)). Even more dramatically, the explicit answerer predicts a uniform distribution over responses to the 'animal?' question. However, the empirical distribution was significantly different from uniform. Thus, the pragmatic answerer is necessary to account for these data.

These data provide strong evidence for a pragmatic answerer, but are more equivocal with respect to the explicit and pragmatic questioner. Because the two models did not make significantly different predictions for this experiment (and both work quite well), we ran a follow-up study on a special case of the guessing-game paradigm in which the explicit and pragmatic questioners make different predictions.

Exp. 2: A Critical Test of Questioner Models

Participants. We recruited 50 participants to participate only in the questioner scenario of the guessing game presented above. Ten participants were excluded on the basis of having a non-English native language, or reporting confusion about the instructions.

Stimuli & Procedure. The procedure was the same as before with some changes to the stimuli. The world space $\mathcal{W}$ consisted of possible assignments of the three pets to three gates. The possible goals $\mathcal{G}$ were the dalmatian and poodle (not the cat). The possible questions $\mathcal{G}$ were 'dalmatian?' or 'cat?'. The possible answers $\mathcal{A}$ were the three gates. Each participant was given the two goals in a random order

Results. When the goal was to find the dalmatian, participants were significantly more likely to ask about the dalmatian than the cat, $\chi^2(1) = 12, p < 0.001$. When the goal was to find the poodle, participants were marginally more likely to ask about the cat than the dalmatian, $\chi^2(1) = 3.6, p = 0.058$. When looking only at the first of the two trials, the dalmatian result held, $\chi^2(1) = 14.4, p < 0.001$, but participants' preference for asking about the cat disappeared, $\chi^2(1) = 0.07, p = 0.79$. These results are shown in Fig. **??**.

Model comparison. The explicit questioner predicts that participants should have no preference for a question given the 'poodle' goal, since an explicit answerer would be equally unlikely to give the desired answer for both. The pragmatic questioner model, however, predicts that participants should prefer to ask about the cat. This is because the (internal) pragmatic answerer would reason that if the questioner was interested in the dalmatian, they would ask about the dalmatian; if they didn't, they must be interested in the other possible goal.

It is again unclear which questioner model is best. Overall, the response distribution matches the predictions of the pragmatic model: questioners prefer to ask about the cat. However, participants don't show this behavior if we look at only the first trial. This could be due to a number of reasons. Interestingly, the pragmatic model predicts a more explicit-like response distribution if the questioner does not take into account the constraint on possible goals: if participants thought the poodle was the only goal (counter to the instructions), then asking about the dog would be consistent with the pragmatic model as well. It is possible that participants only fully-processed the alternative (dalmatian) goal if they had first done the trial where that was the goal.[3]

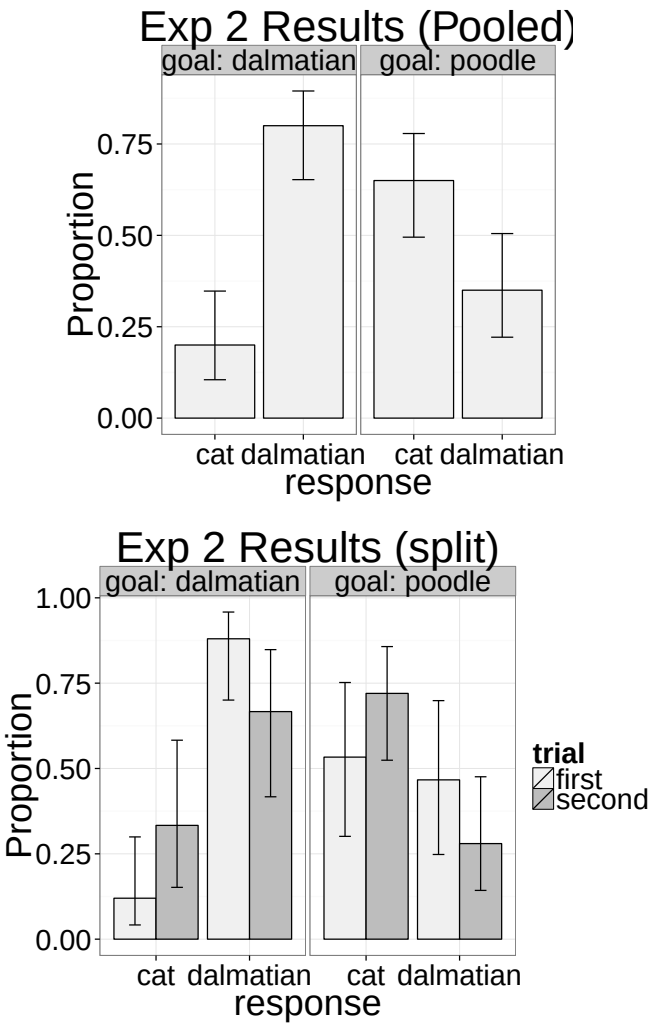

Figure 6. The overall response distribution in Exp. 2 (left) and the same distribution split into first- and second-trial data (right).

Discussion. Experiment 1 established that the pragmatic answerer model is necessary to account for answerer behavior in our simple guessing-game task, complementing results from our computational experiments. Additionally, it established that the explicit and pragmatic questioner models could capture qualitatively important fea-

[3]This pattern was also replicated with a simpler set of stimuli containing red and blue squares and circles.

tures of the questioner data, such as the systematic preference for under-informative questions when the explicit label for an object was unavailable. However, both experiments 1 and 2 failed to distinguish between the explicit and pragmatic questioner models. In experiment 1, they made indistinguishable predictions; in experiment 2, the response data were confounded with participants' beliefs about goal alternatives.

Certain aspects of the task are subject to other methodological concerns. For instance, participants only gave hypothetical judgements about what they *would* say given different goals or questions instead of playing out the full game, and participants did not believe that they were playing with a real partner. These issues contributed to overall confusion about the task, and potentially reflected abstract logic-based reasoning instead of the social and linguistic intuitions we intended to test.

In the following two experiments, we modified the task in two ways to address these concerns. In experiment 3, we replicated experiment 1 in a real-time, multi-player environment where participants are assigned to either the "questioner" role or the "answerer" role and interact directly with one another as they play the full game. In experiment 4, we additionally expanded the set of items beyond the single animal hierarchy used in the previous experiments. This item set includes four different domains (animals, plants, places, and artifacts), crossed with three different hierarchy structures eliciting a larger range of predictions from our models. One of these hierarchy structures was specifically designed to provide a critical test for distinguishing the explicit and pragmatic questioner models. Unlike experiment 2, this condition did not require participants to reason about unintuitive question alternatives that do not pick out any existing items in the world.

Exp. 3: Interactive Questions and Answers

Participants. We recruited 50 participants from Amazon's Mechanical Turk to participate in this task: 25 were assigned to the questioner role and 25 to the answerer role, yielding 25 complete games.

Stimuli & Procedure. The world space $\mathcal{W}$, goal space $\mathcal{G}$, question space $\mathcal{Q}$, and answer space $\mathcal{A}$ were the same as in experiment 1 (see Figure **??**). The procedure was modified to accommodate real-time player-to-player interaction following ? (?). All players passed a short quiz on the game instructions and were immediately redirected to the game interface: the first player to join was assigned to be a "questioner" (which we called a "guesser" in the cover story) and told to wait until a second player was available. Once another player joined, they were assigned to be an "answerer" (or "helper") and the game began.

The questioner and answerer interfaces are displayed in Figure **??**. Chat messages were printed on the left side of the screen, and players used the right side as a workspace to view goals, ask questions, and respond with answers. At the beginning of each trial, the wheel on the questioner's screen (Figure **??**, top) would spin and select one of the four goals. The questioner then clicked and dragged words onto the line in the "Question box" to ask a question to help them find this goal. The answerer saw these words being dragged in real time. Once the questioner clicked the 'send' button, the resulting question appeared in the chat log and control was passed to the answerer, who clicked on one of the four gates to send a response containing the location of the chosen object. Finally, the questioner was asked to guess which gate they believed the goal object was behind.

Each participant provided responses for four trials, where object locations were shuffled and a new goal was randomly selected for each trial. Thus, questioners could be given the same goal on multiple trials, preventing 'process of elimination' reasoning about what the goal may be on the part of the answerer.

Results. Results for the questioner role are shown alongside model predictions in Fig. **??** (left). We find that questioners systematically

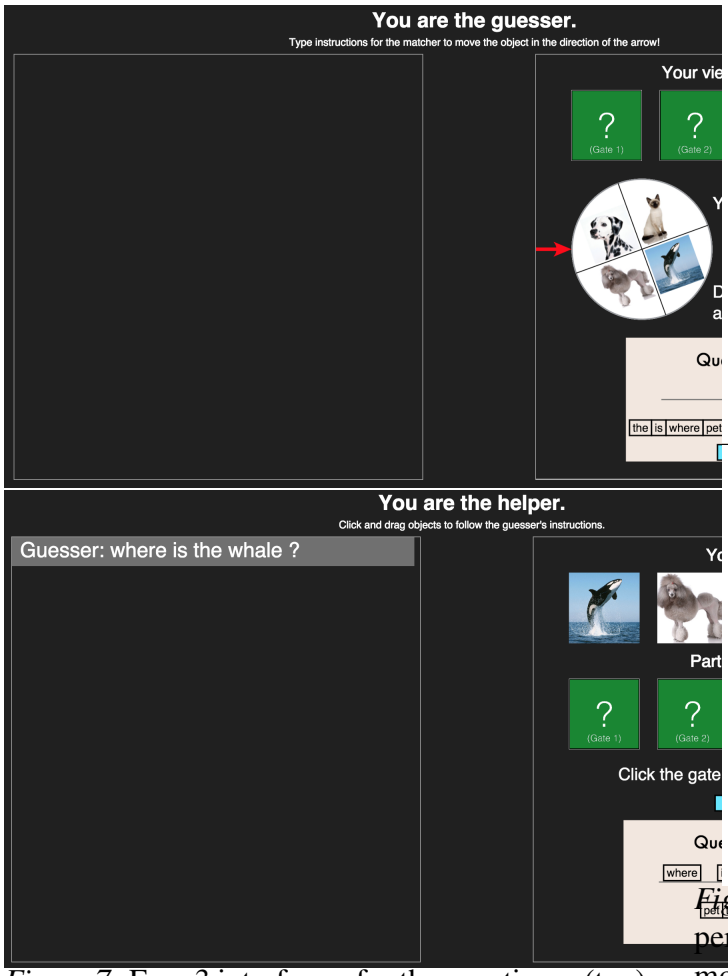

Figure 7. Exp. 3 interfaces, for the questioner (top) and answerer (bottom).

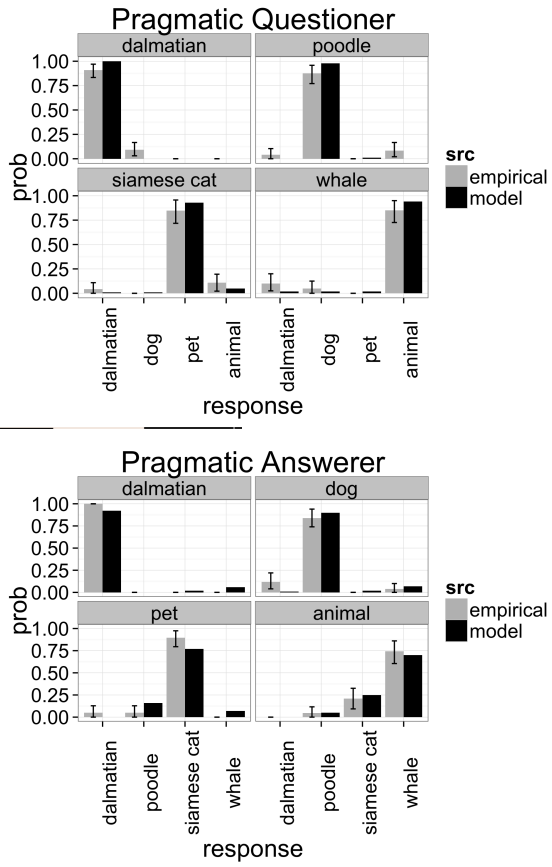

Figure 8. Exp. 3 results and model fits, for the best-performing questioner (left) and answerer (right) models. Error bars are 95% bootstrapped confidence intervals.

prefer to ask different questions given different goals, even as those questions become more indirect. χ^2 tests over each of the four response distributions show a significant divergence from uniform. Questioners preferentially ask about the 'dalmatian' given the dalmatian goal, $\chi^2(3) = 77, p < .001$, about the 'dog' given the poodle goal, $\chi^2(3) = 50, p < .001$, about the 'pet' given the cat goal, $\chi^2(3) = 47, p < .001$, and about the 'animal' when given the whale goal, $\chi^2(3) = 39, p < .001$. Note that there was a high level of agreement between participants: responses were not particularly graded.

Results for the answerer role are shown in Fig. **??** (right). Answerers are highly sensitive to the constraints of the questioner, giving information about the dalmatian when asked about a 'dalmatian', $\chi^2(3) = 102, p < .001$, about the poodle when asked about a 'dog', $\chi^2(3) = 47, p < .001$, about the cat when asked about a 'pet', $\chi^2(3) = 45, p < .001$, and about the whale when asked about an 'animal', $\chi^2(3) = 31, p < .001$. In the next section, we compare these results to the predictions of our family of models (Fig. **??**).

Model comparison. Model comparison was conducted in the same way as in Experiment 1: a single optimality parameter, which applied to all agents, was fit for each of the six models to maximize correlation with the data.

We can again rule out both the literal answerer and literal questioner as they predict a uniform distribution of responses over the four questions

and answers. The two remaining questioner models again make roughly the same predictions for this task: we found a model-data correlation of $r = 0.971$ for the explicit questioner and correlation of $r = 0.996$ for the pragmatic questioner. The difference between these correlations is statistically significant, accounting for their shared dependence on the empirical data (Zou's confidence interval = $[-0.079, -0.009]$). However, they make nearly identical qualitative predictions; the pragmatic questioner model's predictions for each response distribution are shown in Fig. **??** (left).

The pragmatic answerer provides a much better fit to the data than the explicit answerer: we find a model-data correlation of $r = 0.7$ for the explicit answerer and $r = 0.99$ for the pragmatic answerer. Taking into account the fact that these correlations are dependent and overlapping on the same empirical data, we find that the pragmatic answerer correlation is significantly larger than the explicit answerer correlation (Zou's confidence interval = $[-0.676, -0.107]$).

Discussion. We replicated the results of experiment 1 in a real-time, interactive setting. Again, we found that both explicit and pragmatic questioner models provide an excellent fit to questioner behavior, and that the pragmatic answerer accounts for the data significantly better than the explicit answerer both quantitatively and qualitatively. In addition, the full, interactive game was designed to be more natural and less confusing to participants than the drop-down menu design from experiments 1 and 2. Because players received constant feedback from their partner, this task was framed as inherently social, and because answerers watched as questioners moved words to form questions, there was a convincing mechanism for people to believe they were playing with another human. This addresses some concerns raised with experiments 1 and 2.

Note that so far we have used the same animal hierarchy for the stimulus set in all our experiments, providing only 16 points of comparison between our models and empirical data. Further-

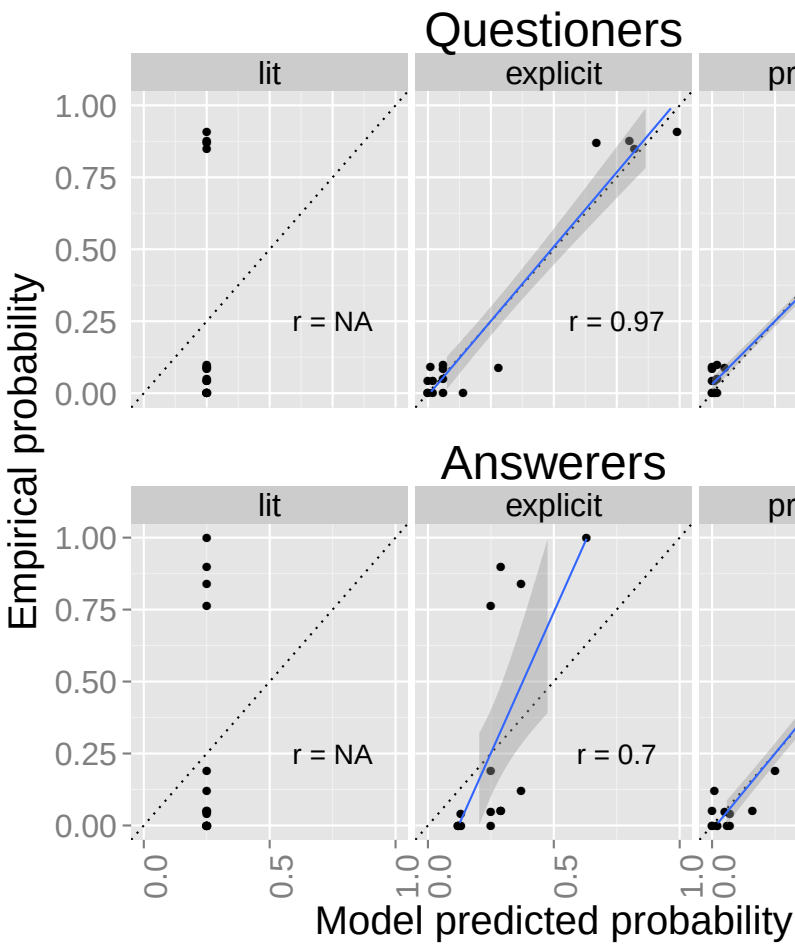

Figure 9. Full space of models, and their correlations with the data from Exp. 1. Questioner models in the first row reason about the answerers directly below them, and the pragmatic answerer reasons about the explicit questioner.

more, there exists a heuristic strategy for selecting questions given goals in the hierarchy structure we have been using which produces the same pattern of responses without requiring any social inference. Suppose questioners saw their goal on a given trial and ruled out labels that do not apply (e.g. a 'cat' is neither a 'dalmatian' nor a 'dog'), then picked the most specific of the remaining labels ('pet' picks out a smaller set of objects than 'animal'). Whether through use of this heuristic strategy or pragmatic inference (as our model sug-

gests), questioners strongly converged on a single mapping from goals to question utterances. Because response probabilities are clumped at the ends of the scale, our fits to the questioner model are not particularly meaningful: we're fitting endpoints.

In experiment 4, we test the generality of our model by expanding the stimulus set to encompass multiple stimulus domains and multiple hierarchy structures designed to elicit graded judgements. This addresses the potential concern that the behavioral patterns we have been modeling are specific to the set of four animals or the tree-like hierarchy in which they were embedded. We also took care to include one simple but critical hierarchy structure where (1) the explicit and pragmatic questioner models make different predictions and (2) the heuristic strategy cannot mimic these predictions.

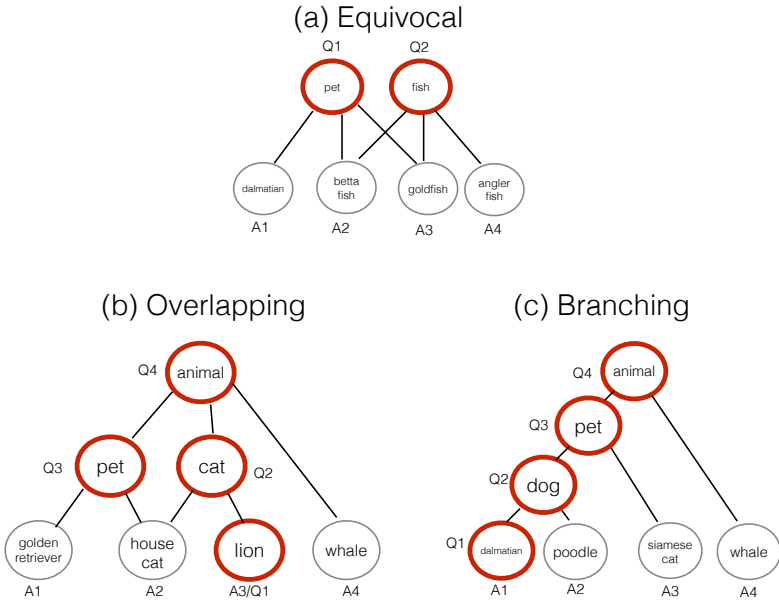

Figure 10. Example of each type of hierarchy used in Experiment 4. The bottom row served as the answer space, and the labels in the questioner space highlighted in red.

Exp. 4: Generalizing Predictions

Participants. We recruited 199 participants from Amazon's Mechanical Turk to participate in this task. Fifty participants were excluded due to a server crash that terminated the task before completion. Two additional participants were excluded because they were not non-native English speakers. This left 74 unique completed games.

Stimuli & Procedure. A set of twelve items were created by crossing four domains (animals, plants, places, and artifacts) with three hierarchy structures ("branching", "overlapping", and "equivocal"; see Figure **??**). For each item, there was a set of four goal objects in $\mathcal{G}$, which appeared on the wheel for the questioners, four question labels in $\mathcal{Q}$ that the questioner could use to ask about their assigned goal, and four answers in $\mathcal{A}$ corresponding to the four items in $\mathcal{G}$ that appear on the wheel.[4]

The procedure was the same as Experiment 3 with two major modifications. First, we addressed an asymmetry in the task: the answerer watched the questioner move words, but the questioner only saw template text responses in response. In pilot work, we found that answerers were more likely to believe they were playing with a human partner than questioners. To bring questioner and answerer beliefs closer together, we changed the mechanism by which the answerer responds. Instead of clicking on a gate and then clicking the 'send' button, we supplied a "Reveal Box." When the answerer was ready to reveal a gate, they would simply click and drag the object they wanted to reveal into this box. The questioner watched as the outline of the gate moved, in real time, and saw the image as soon as the answerer dropped it in the box.

Second, we noticed that some players began exploiting the freedom of the "question box" drag-and-drop procedure after playing several rounds. On difficult items, especially those in the equivocal hierarchy where our model predicts no preferred question for certain goals, players would use the

[4] A document containing these mappings for all items, as well as their hierarchical relationships, is available online at `https://github.com/hawkrobe/Q_and_A/blob/master/Mult`

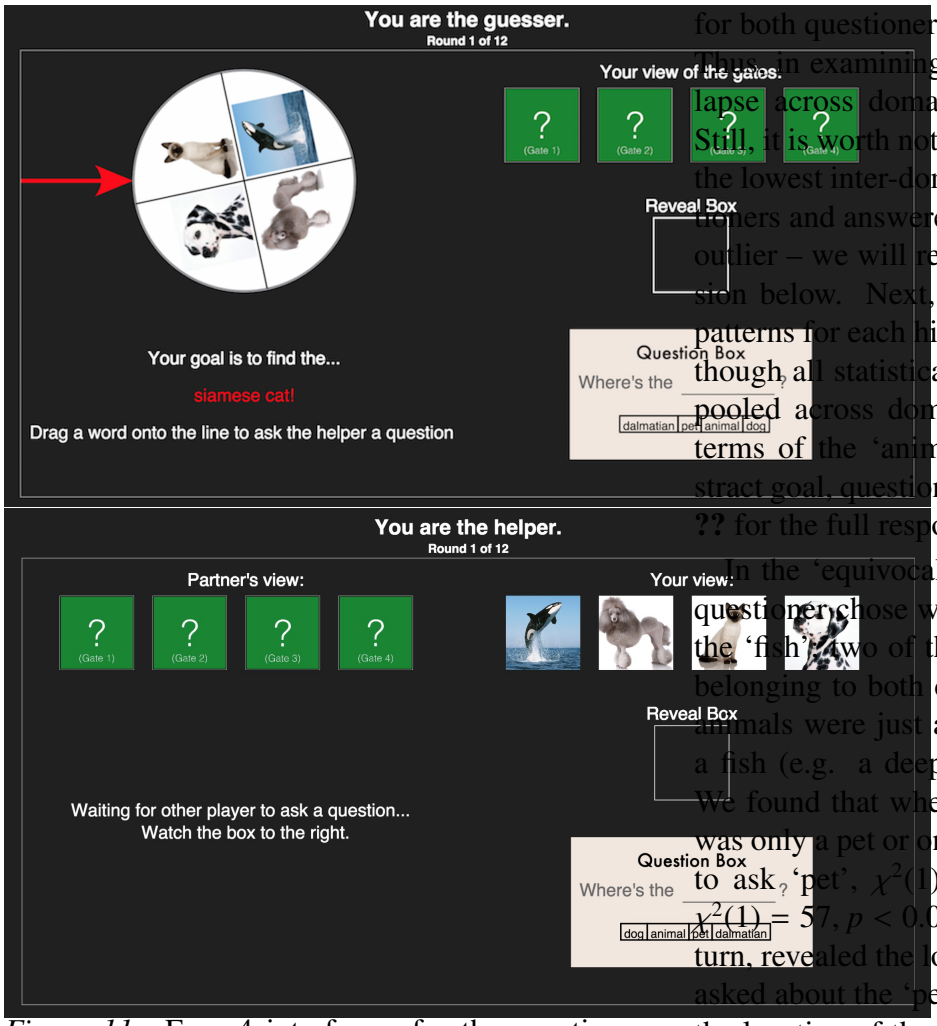

Figure 11. Exp. 4 interfaces, for the questioner (top) and answerer (bottom).

words to form non-grammatical signals. This is interesting behavior in its own right, but we are interested in testing models which operate over a well-defined set of question utterances held in common ground. We therefore pre-set a frame "Where is the . . . " for the question box and allowed participants to drag one and only one of the question labels into the blank.

The resulting questioner and answerer interfaces are displayed in Figure **??**. Each participant provided one response for each of the twelve items, presented in random order.

Results. First, we note that there was high correlation in response probabilities across domains for both questioners and answerers (see Table **??**). Thus in examining the results below we will collapse across domains for simplicity of analysis. Still, it is worth noting that the "place" domain has the lowest inter-domain correlations for both questioners and answerers, indicating that it may be an outlier – we will return to this point in the discussion below. Next, we step through the response patterns for each hierarchy type. For concreteness, though all statistical tests were conducted on data pooled across domains, we report our results in terms of the 'animal' domain instead of the abstract goal, question, and answer labels (see Figure **??** for the full response distributions).

In the 'equivocal' condition (Figure **??**(a)), the questioner chose whether to ask about the 'pet' or the 'fish'. Two of the goal animals were 'pet fish' belonging to both categories, while the other two animals were just a pet (e.g. a dalmatian) or just a fish (e.g. a deep-sea angler fish), respectively. We found that when trying to find the object that was only a pet or only a fish, participants preferred to ask 'pet', $\chi^2(1) = 67, p < 0.001$, or 'fish,' $\chi^2(1) = 57, p < 0.001$, respectively. Answerers, in turn, revealed the location of the 'dalmatian' when asked about the 'pet', $\chi^2(3) = 238, p < 0.001$, and the location of the 'angler fish' when asked about the 'fish,' $\chi^2(3) = 142, p < 0.001$.

For the two objects belonging to both categories, where we expected questioners to have no preference, we found strong variability across domains – in the 'animal' domain, for example, 85% of questioners asked about the 'fish' when given one of the 'pet fish' goals, compared to only 15% asking about the 'pet,' $\chi^2(1) = 15, p < 0.001$. In the 'artifacts' domain, on the other hand, questioners had no preference between asking about the 'seat' or 'metal thing' when the objects (metal chairs) fell into both categories, $\chi^2(1) = 0.4, p = 0.52$. This suggests that labels have differential fitness when applied to different objects, and participants are relying on more knowledge than the purely structural connections captured by the hierarchy.

In the 'overlapping' condition (Figure **??**(b)),

questioners preferentially asked about the 'lion' when looking for the lion, $\chi^2(3) = 205, p < 0.001$; about the 'cat' when looking for the Siamese cat, $\chi^2(3) = 63, p < 0.001$, even though the lion is also a cat; about the 'pet' when looking for the dalmatian, $\chi^2(3) = 232, p < 0.001$, even though the Siamese cat is also a pet; and about the 'animal' when looking for the whale, $\chi^2(3) = 213, p < 0.001$. Answerers preferentially revealed the lion when asked about the 'lion,' $\chi^2(3) = 229, p < 0.001$, the Siamese cat when asked about the 'cat,' $\chi^2(3) = 59, p < 0.001$, the dalmatian when asked about the 'pet,' $\chi^2(3) = 120, p < 0.001$, and the whale when asked about the 'animal,' $\chi^2(3) = 187, p < 0.001$.

In the 'branching' hierarchy (Figure **??**(c)), we replicated our findings from Experiments 1 and 3 with a broader set of items: questioners strongly prefer to ask about the 'dalmatian' when trying to find the dalmatian, $\chi^2(3) = 190, p < 0.001$, about the 'dog' when trying to find the poodle, $\chi^2(3) = 152, p < 0.001$, about the 'pet' when trying to find the siamese cat, $\chi^2(3) = 168, p < 0.001$, and about the 'animal' when trying to find the whale, $\chi^2(3) = 210, p < 0.001$. Answerers behave identically to previous experiments, $p < 0.001$.

Model comparison. Model comparison was conducted in the same way as in Experiment 1 and 3. However, to avoid overfitting and to reduce the number of free parameters that would result from separately fitting rationality parameters for each item, we pool together data from all hierarchy structures and domains and fit the single rationality parameter for each model to jointly optimize model-data correlations across all items in this pooled dataset.

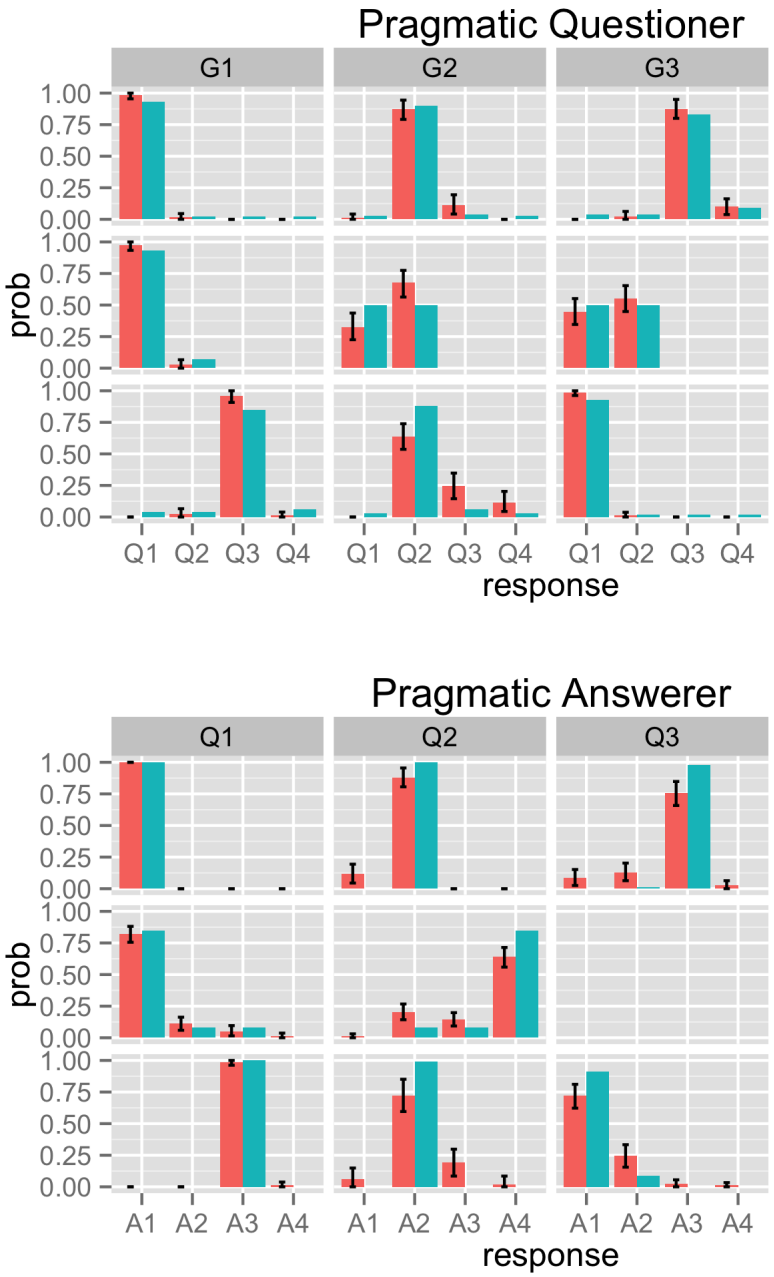

Figure 12. Exp. 4 results and model fits, for the best-performing questioner (left) and answerer (right) models, collapsing over the different domains. Error bars represent bootstrapped 95% confidence intervals.

	Questioners				Answerers			
	animal	place	plant	artifact	animal	place	plant	artifact
animal	1.00	0.91	0.94	0.94	1.00	0.78	0.92	0.97
place	0.91	1.00	0.96	0.95	0.78	1.00	0.78	0.78
plant	0.94	0.96	1.00	0.97	0.92	0.78	1.00	0.91
artifact	0.94	0.95	0.97	1.00	0.97	0.78	0.91	1.00

Table 1
Inter-domain correlations in Experiment 4

These model-data fits, broken down by model

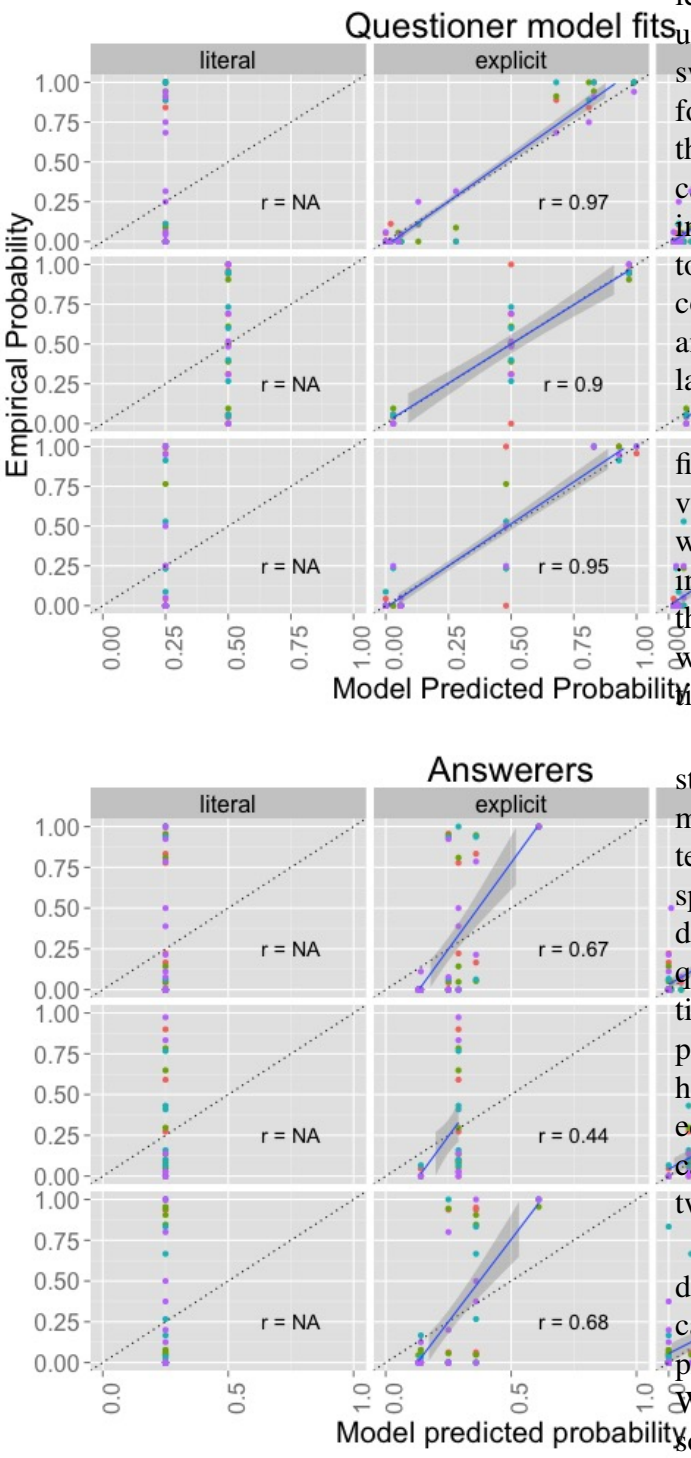

Figure 13. Full space of models, and their correlations with the data from Exp. 4, broken down by item and type.

level and hierarchy structure, are displayed in Figure **??**. We can immediately rule out the literal answerer and literal questioner, which predict a uniform distribution of responses. We also find that the explicit answerer models fit the data significantly worse than the pragmatic answerer model in all hierarchy structures, with $r = 0.67$ compared to $r = 0.96$ in the 'branching' hierarchy, $r = 0.44$ compared to $r = 0.92$ in the 'equivocal' hierarchy, and $r = 0.68$ compared with $r = 0.86$ in the 'overlapping' hierarchy, respectively.

Turning to the questioner predictions, we again find that both explicit and pragmatic models provide excellent but essentially indistinguishable fits, with $r = 0.97$ compared with 0.98 in the 'branching' hierarchy, $r = 0.90$ compared with $r = 0.90$ in the 'equivocal hierarchy', and $r = 0.95$ compared with 0.94 in the 'overlapping' hierarchy, respectively.

However, because we designed the overlapping structure to provide a critical test for the questioner models, it is not the overall correlation we are interested in, it is the particular distribution of responses to "G2" (the "house cat" in the animal domain). Since "house cat" is not available as a question, the explicit model (as well as the heuristic strategy discussed above) predicts that the two parents, "pet" and "cat" will be equally likely. This happens because the explicit answerer would be equally likely to reveal the location of the house cat in response to "pet" and "cat," which each have two children.

The pragmatic questioner model, however, predicts that "cat" will be preferred over "pet", because the other child of "cat" – the lion – is pragmatically blocked by the pragmatic answerer. When the pragmatic answerer hears "cat" they reason that if the questioner's underlying goal were the "lion" they would have *said* lion; because they didn't, they must mean the other cat. Thus, we have a pair of sharply distinguishable predictions: the explicit model predicts that "pet" and "cat" will be equally likely and the pragmatic model predicts an asymmetry where "cat" is the preferred label.

In Figure **??**, we zoom in on this critical condition and display both models' predictions next to the empirical data, with bootstrapped 95% confidence intervals. For these analyses, we excluded the 'places' domain both because it was flagged as an outlier early in the analysis and because our stimulus design led to a confound in the 'overlapping' condition, which we discuss further below. We find that the pragmatic model makes the correct qualitative prediction – the mean probability of responding "Q2" ($p_{Q2} = 0.77, n = 52$, 95% bootstrapped CI $= [0.65, 0.88]$) in this condition is significantly different from the mean probability of responding "Q3" ($p_{Q3} = 0.15, n = 52$, 95% bootstrapped CI $= [0.06, 0.25]$), as predicted by the pragmatic model. We also see that it makes a relatively good quantitative prediction, with the predicted probability lying within the 95% confidence interval of the empirical estimate for both the "Q2" and "Q3" responses. Recall that a single free parameter α for each model was fit to maximize correlations for the entire (pooled) dataset.

Discussion. We have again replicated the results of Experiments 1 and 3, and shown that these results generalize to a wider range of domains and hierarchy structures. Additionally, we were finally able to distinguish between the pragmatic and explicit questioner models, finding that the pragmatic model is necessary to account for some critical aspects of the data. The "place" domain that we flagged as a potential outlier earlier in the results is the only domain that does not show the pattern of responses predicted by the pragmatic questioner model. This is likely due to the choice of images displayed to participants: the two "parent" nodes of the critical goal are "bar" and "restaurant," and we represented the intersection of these two categories as a hotel lobby. However, the image we chose placed emphasis in the bar in the hotel, which may have biased participants toward the "bar" response instead of the "restaurant response" predicted by the pragmatic model. Including it does not change the statistical results, but we do not believe it is representative of questioner behavior.

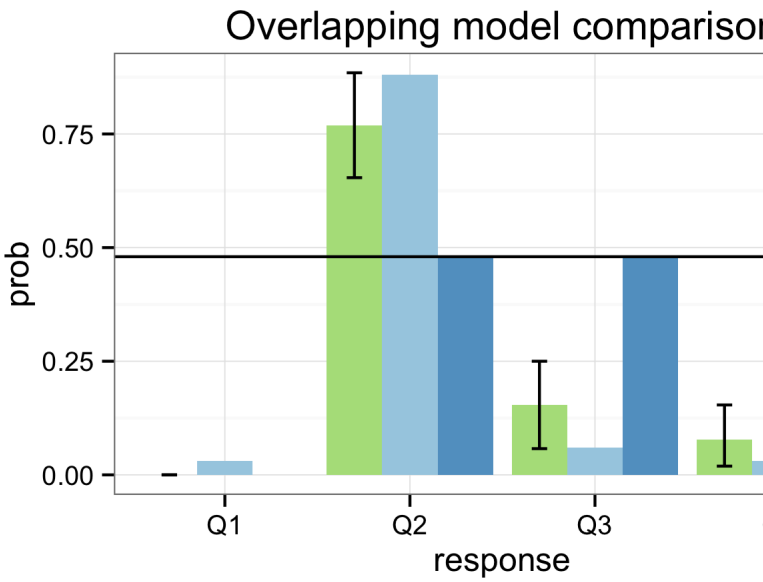

Figure 14. Full space of models, and their correlations with the data from Exp. 4, broken down by item and type.

The next paragraph would naturally lead into another experiment where we elicit label priors and re-run our model with label fitness included – I think I'd be happier with the paper if we had this...

We also find that while overall fits are good, all our models fail to capture graded responses across different domains. This is especially striking in the 'equivocal' condition where all our models predict a symmetry in questioner behavior for the two goals where the questions options are equally uninformative. Instead, we found a high degree of variability across domains. To some extent, this is expected from a model with only a single parameter fit for all domains simultaneously. We could have fit separate rationality parameters for each domain to capture domain-level differences. However, we expect that this variability can be accounted for in a more principled way by a prior over label fitness. In our data, 'fish' was considered a better

label for a picture of a pet goldfish than 'pet' by 85% of participants, even though both are technically true. When there is no pragmatic reason to favor one over the other, participants fall back on their label prior. If we ran a separate study to elicit this prior, we could simply plug the empirical priors into our model and naturally account for the variability across items without introducing additional free parameters. Alternatively, we could use Bayesian data analysis to infer these priors from question and answer response data.

General discussion

Perhaps the most important formal advance of the models considered here is to move the Rational Speech Act framework beyond interpretation of single utterances (in context), to consider the dynamics of simple dialogs (albeit consisting of a single question and its answer). Doing so requires replacing the immediate motive to convey true information with the more distant motive to provoke useful information from one's interlocutor. On the answerer side, sophisticated inference was required to account for the implicit interests of the questioner. This provides a useful connection to current game-theoretic and decision-theoretic models (?, ?, ?), which also emphasize the importance of goals and speaker beliefs in communication but emphasize less the complex interplay of inference between questioner and answerer.

We have presented evidence that answerer behavior is best described by a pragmatic model that *does* reason about questioner intentions, using the question utterance as a signal. Our study provides a next step for previous intention-based linguistic accounts by concretely specifying how an answerer may make such inferences via Bayesian conditioning. The superiority of pragmatic answerer predictions over the other answerer models was robust across all experiments.

Questioner behavior in Exp. 2, however, seemed to be much more dependent on experience. In another version of Exp. 1, we did not emphasize certain aspects of the game in the instructions, such as the fact that the answerer knows about the restricted answer set, which might prompt perspective-taking. Our data in this pilot experiment appeared to contain a mixture of explicit and pragmatic answerers and questioners (though other confounds were present in this version). We found the interactive, multi-player version of the task, used in Exp. 3 and Exp. 4 to be more robust to these minor variations. We also found that at least in certain scenarios like the overlapping hierarchy condition of Exp .4, questioners systematically relied on higher-order pragmatic reasoning about what inferences an answerer would make about their own underlying goals when deciding what question to ask. Note that this behavior also could not be explained by the heuristic strategy raised in the earlier experiments: if questioners just ruled out labels that did not apply (e.g. "lion"), they would have no mechanism for deciding between the two equally-good parent labels ("pet" and "cat") to pick out their goal. It will ultimately be important to explore the mixture of explicit- and pragmatic-questioning across an even larger range of situations: these issues may be a product of our artificial game paradigm, or they may be reflective of real tendencies in language use, raising novel questions about audience design in question-answer behavior.

While the artificiality of our question-answer game may distance the behavior of participants from the natural use of language, there are also some benefits to this design. In particular, it is easy in this setting to control the exact space of questions, goals, and answers. While the restrictions on question space may seem peculiar, it is directly motivated by conversational scenarios in everyday usage which feature restrictions on the set of things one can ask about, due to politeness, salience, time cost, and other factors. In future work, we will explore the extent to which the proposed model can scale up to extended dialogues, and other more naturalistic language settings. To deal with dialogues lasting longer than a single exchange, for instance, we must specify the way in which the contribu-

tions of questioner and answerer affect the *context* in which later utterances operate.

Humans are experts at inferring the intentions of other agents from their actions (?, ?). Given simple motion cues, for example, we are able to reliably discern high-level goals such as chasing, fighting, courting, or playing (?, ?, ?). Experiments in psycholinguistics have shown that this expertise extends to speech acts. Behind every question lies a goal or intention. This could be an intention to obtain an explicit piece of information ("Where can I get a newspaper?"), signal some common ground ("Did you see the game last night?"), test the answerer's knowledge ("If I add these numbers together, what do I get?"), politely request the audience to take some action ("Could you pass the salt?"), or just to make open-ended small talk ("How was your weekend?"). These wildly different intentions seem to warrant different kinds of answers. By formalizing the computational process by which answerers infer these different intentions, our model framework provides a unifying way to accommodate this diversity.