Questions and answers in dialogue

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

What makes a question useful? What makes an answer helpful? Asking questions is one of our most efficient and reliable means of learning about the world. Yet we do not often pose these questions to an impartial oracle; we ask other agents, in dialogue. In this paper, we reconcile formal models of optimal question-asking with classic psycholinguistic effects of social context by extending recent probabilistic cognitive models of goal-relevant communication to dialogue. Agents hear an utterance, update their beliefs through social reasoning, and select an utterance in response. The utility of producing a question in this framework is the expected reduction in uncertainty about the aspects of the world relevant to the speaker's goal upon hearing an answer from their partner. Critically, this suggests that the question itself can be a signal to the speaker's underlying goals, and that a helpful answerer should be informative with respect to inferred goals, beyond the literal meaning of the question. We compare models of both the questioner and answerer within this framework based on three pieces of evidence: First, we account for classic effects in psycholinguistics showing that the same question can yield different answers depending on the context. Second, we run a real-time, multi-player communication game to jointly test predictions made by the questioner and answerer components of our model. Third, we show that the salience of goals affects both asking and answering questions. We find that sophisticated pragmatic reasoning in dialogue is needed to account for critical aspects of the data.

DRAFT 2/13//2018: THIS PAPER HAS NOT BEEN PEER REVIEWED. PLEASE DO NOT COPY OR CITE WITHOUT AUTHOR'S PERMISSION

Keywords: pragmatics, language, computational modeling, active learning, questions, answers

The pragmatics of asking and answering

Asking questions is one of our most valuable methods of gathering information and learning from others. At a very young age, children can strategically select *who* to ask (Legare, Mills, Souza, Plummer, & Yasskin, 2013) and *what* to ask (Chouinard, 2007; Ruggeri & Lombrozo, 2015) in order to solve problems and test lay theories (Callanan & Oakes, 1992). Robots that ask questions to clarify commands (Deits et al., 2013) or obtain assistance with a task (Fong, Thorpe, & Baur, 2003) make for better collaborators. In tutoring scenarios, students who ask better questions

This report is based in part on work presented at the 37th Conference of the Cognitive Science Society. Correspondence concerning this article should be addressed to Robert X.D. Hawkins, e-mail: rxdh@stanford.edu

tend to be more successful (Graesser & Person, 1994), and scientists who design their studies to ask better questions can obtain much more informative results (Clark & Schober, 1992; Myung & Pitt, 2009).

A question is only as useful as its answer, however, and not all answers are equal. How does an answerer select from the broad range of possible responses? This is straightforward in some cases. For direct, factual questions, an answerer can use the explicit meaning of the question, which specifies a particular gap in the questioner's knowledge, and respond with the requested piece of information (if it is known):

(1) Q: Who was the 16th president of the U.S.?
A: Abraham Lincoln.

This response strategy has been implemented in sophisticated semantic parsers that extract the logical form of the query and efficiently search a database for an entity that satisfies it (e.g. Berant, Chou, Frostig, & Liang, 2013).

Many questions in everyday conversation, however, are not so straightforward to answer. For a class of questions called indirect speech acts, the questioner typically will not be satisfied with a simple yes/no answer; they expect the answerer to go *beyond* the literal meaning to address the questioner's true interests (Clark, 1979):

(2) Q: Do you know the time?
A: It's a quarter past 3.

Indeed, answerers often go beyond the literal meaning even for *direct* questions. A recent corpus study found that 14% of responses to direct yes/no questions used a full statement instead of a simple "yes" or "no" (de Marneffe, Grimm, & Potts, 2009):

(3) Q: Is it in Dallas?
A: Uh, it's in Lewisville.

In many contexts, radically underspecified questions are the rule rather than the exception. Teachers, doctors, technical troubleshooters, concierges, and others in the service industry are regularly faced with the challenge of responding helpfully even when the questioner may not be clear about precisely what information they need:

(4) Q: What's wrong with my answer to #1?
A: Well, let me remind you how to compute a derivative...

What do these exchanges have in common? Q chooses a question that manages to send a signal to A about her intentions, while sometimes being unable or unwilling to specify exactly what those intentions are. They may be impolite or embarrassing, they may be too long or costly to fully explain, or she may just not know enough about the topic she's interested in to articulate them.

A, in turn, reasons under uncertainty beyond the overt question and provides an answer that addresses Q's true interests. Depending on the circumstances, he can adjust his response to be over- or under-informative with respect to the direct question, or to address a different question altogether. This subtle context-sensitive interplay between a questioner choosing a question to ask and an answerer choosing a response raises two specific questions for formal models of linguistic behavior: What makes a question useful? And what makes an answer helpful?

We suggest, following Van Rooy (2003) and other recent approaches (see Coenen, Nelson, & Gureckis, 2017, for a thorough review), that the value of a natural-language question is the extent to which it can be expected to elicit useful information. More specifically, however, the value of a question is the expected information gained relevant to the questioner's interests, given the set of likely answers it may provoke *from another agent*. The value of an answer, then, is the extent to which it resolves questioner uncertainty over the *goal-relevant* aspect of the world.

This paper makes two key theoretical contributions. First, while psychologists have learned a lot about questioner behavior and answerer behavior by studying the two processes in isolation, we

argue that there are significant theoretical benefits in viewing them as deeply intertwined *social* inferences. Second, we formalize this view in a probabilistic model of pragmatic language understanding, bridging the gap between the simple queries used in psychological research on optimal information search and the natural language questions studied by linguistics. This is the first time that such pragmatic language models have been extended beyond single utterances and hence provides a first step toward modeling longer dialogues.

The rest of this paper is structured as follows. First, we review previous experimental and formal modeling efforts to situate question-asking and question-answering in a social context. We then lay out the details of our computational framework, formalizing different hypotheses in distinct questioner and answerer models. Finally, we evaluate these models in two ways: through a set of computational experiments capturing three classic *answerer* sensitivity effects, and through three novel behavioral experiments addressing a longstanding paucity of data on social reasoning in *questioner* behavior. These behavioral experiments develop a novel cooperative communication paradigm where different questioner and answerer models can be rigorously distinguished through both qualitative predictions and quantitative model comparison. In particular, this paradigm addresses the experimental challenge of inducing uncertainty for the answerer over the questioner's private goal. By showing how the model handles graded concept knowledge, structured goals, and multi-round dialogue, we demonstrate how our core computational framework can be elegantly composed with different cognitive components and representations to capture a wide variety of phenomena.

Empirical background

A number of psycholinguistic studies have provided evidence that answerers are both sensitive to a questioner's goals and attempt to be informative with respect to those goals. For instance, in Clark's (1979) classic study, researchers called liquor merchants and asked "Does a fifth of Jim Beam cost more than $5?" Before asking the question, they provided some context for their call by saying either "I want to buy some bourbon" (the *uninformative* condition) or "I've got $5 to spend" (the *five dollar* condition). These contexts were designed to signal different speaker goals. In the uninformative condition, the goal is simply to buy whiskey, hence the exact price would be useful information; in the five dollar condition, the goal was literally to find out whether or not the speaker could *afford* the whiskey, so a yes/no answer would suffice. If merchants inferred these goals from the context signal and responded with respect to these goals, we would expect different types of answers in the two conditions. Indeed, merchants gave a literal yes/no answer more often in the latter condition than the former, where an exact price was more common.

Goals can also be inferred from non-linguistic environmental cues. For instance, Der Henst, Carles, and Sperber (2002) investigated answers to questions like "Do you have the time?" that permit answers with different degrees of approximation to the true time (see also Gibbs Jr & Bryant, 2008). In a baseline condition, participants approached in public typically rounded their answers to the nearest 5 or 10 minute interval even when they were wearing a digital watch. When the experimenter mentioned that they had an appointment at a given time, however, answers became more precise as the true time approached the appointment time, indicating sensitivity to implicit time-related goals like being on time.

Linguists interested in pragmatic accounts of answerer behavior have noted a number of additional scenarios with more nebulous goals. Questions like "where are you?" permit answers at many degrees of specificity: *the United States*, *my apartment*, and *by the big tree* are each perfectly appropriate in some context and highly inappropriate in others (Potts, 2012). Identification questions like "who is X?" can be resolved in many ways (Boër & Lycan, 1975; Gerbrandy, 2000; Aloni,

2005). For example, if an undergraduate asked "Who is Noam Chomsky?" in an introductory course, it would be appropriate to respond "The MIT professor who wrote *Aspects of the Theory of Syntax*" or "The father of modern linguistics." If a potential donor asked the same question at an MIT fundraiser, though, it would be more appropriate to point at Chomsky in the crowd. Similarly, if a child asks "What's that?" while pointing at a common household object, a parent's response will be much different than if one of their adult friends asked the same question. More specific *wh*-questions like "Who passed the examination?" are also context-sensitive: answers can be understood to mean either an *exhaustive* or *selective* list of relevant entities depending on the scenario (Schulz & Van Rooij, 2006).

Taken together, this body of work suggests that *goal-relevance* is a critical feature of question-answering: answerers routinely go beyond the literal meaning of a question to help the questioner achieve their underlying goals. Note that the vast majority of prior work on question-answer behavior has focused on the *answerer*, holding the question constant and investigating the effect of different contexts. However, questioners must also select between many alternative questions in order to achieve their goal and it remains unclear to what degree their choice is affected by social context. Next, we review a set of formal proposals about the questioner's role in dialogue.

Modeling background

An artificial agent capable of flexibly answering questions posed in natural language would be valuable for a diversity of practical applications, from customer service and technical support to health care and financial consulting. It's no surprise, then, that the problem of question-answering has attracted attention from researchers in the artificial intelligence community for decades (Simmons, 1965; Lehnert, 1977; Allen & Perrault, 1980; Green & Carberry, 1994; Mollá & Vicedo, 2007). Some of these classic systems are highly domain-restricted interfaces for databases, such as the kind that could be used to show airline customers flights that meet their criteria, but many built on insights from early work in cognitive science to represent questioner intentions as schemas or plans. For instance, the system designed by Allen and Perrault (1980) uses a set of logical rules to generate a space of possible plans, searches this space using a set of heuristics, identifies obstacles to fulfilling this plan, and takes a set of actions to address the obstacles.

At the same time, linguists have developed formal theories of what questions and answers mean in the first place, mainly focusing on the notion of informativeness. In Groenendijk and Stokhof's (1984) 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 complete 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 (2003) and others (Ginzburg, 1995) have pointed out, this predicts that wh-questions like "Where do they sell Italian newspapers in Amsterdam?" can only be fully resolved by exhaustively mentioning whether or not such a newspaper can be bought at each possible location in the city. Clearly, this is not the case: a single nearby location would suffice. Even theories that allow for 'mention-some' answers cannot account for contextual variation in what counts as a useful answer: if a media executive asked this same question at a corporate meeting, they might legitimately be requesting an exhaustive list.

More recent linguistic theories have tried to fix these problems by introducing some consideration of the questioner's goals. Van Rooy (2003), for instance, formalizes these goals in terms of a decision problem faced by the questioner and assumed to be in common ground. He considers two proposals for how meanings depend on this decision problem, ultimately arguing for the second.

In the first proposal, questions have a context-independent meaning – in the sense that an interrogative sentence always creates the same partition on worlds – but the extent to which an answer is *useful* or *resolving* is determined by the decision problem. In particular, a useful answer under this account is one that maximizes the expected value of the questioner's decision problem, even if it is not a complete semantic answer.

In the second proposal, the meaning of a question is *underspecified* and its interpretation depends on the decision problem. In particular, because the questioner is assumed to be rational, the answerer chooses the interpretation that would maximize the expected value of the answer. We take these proposals as a starting point for the meaning of questions in our models.

A Rational Speech Act model of questions and answers in dialogue

Overview

Broadly speaking, an agent must solve two problems to participate in a dialogue. First, they must *listen*, interpreting the meaning of their partner's utterance and updating their own beliefs accordingly. Then they must *speak*, producing an utterance that informatively addresses the current goal or topic of conversation. At either stage, the utterance may be a statement, a question, or another more exotic kind of speech act.

Both the listening problem and the speaking problem have been addressed separately within the Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Goodman & Frank, 2016), in which language understanding is formalized as recursive Bayesian inference. Pragmatic speaker agents choose utterances to maximize their partner's surprisal about the true state of the world – a basic epistemic goal – and pragmatic listener agents interpret utterances by inverting the speaker model. This basic framework has been used to account for a number of diverse pragmatic phenomena including scalar implicature (Goodman & Stuhlmüller, 2013), generic language (Tessler & Goodman, 2016), and interpretation of context-sensitive adjectives like "tall" or "cheap" (Lassiter & Goodman, 2015).

RSA models have also recently been extended to incorporate the critical notion of *goal-relevance* into the speaker's utility (Roberts, 1996; Wilson & Sperber, 2012). Instead of attempting to be maximally informative about the *full* state of the world, the speaker ignores irrelevant dimensions and only attempts to be informative about the relevant feature or summary statistic. RSA models with relevance have been used to capture non-literal language use like hyperbole (Kao, Wu, Bergen, & Goodman, 2014), metaphor (Kao, Bergen, & Goodman, 2014), or irony (Kao & Goodman, 2015). In the case of hyperbole, for instance, a speaker who says "It took a million years to get a table" may be intending to be informative about the topic of their affective state (e.g. their frustration) rather than the exact time it took to get a table.

Here, we build upon these recent modeling advances in several novel ways. First, by unifying the listening problem and speaking problem in a single rational agent, we provide a principled cognitive connection between linguistic input and output: different kinds of input shift an agent's beliefs in different ways, motivating different kinds of responses.

Additionally, questions have previously posed a challenge for rational models of language use, since questions – as utterances – don't provide direct information about the world to the listener. The crucial motivation to ask questions in the first place, of course, is an asymmetry between speaker and listener knowledge; agents have *private goals* in addition to private knowledge about the true state of the world. While a declarative statement in our framework provides evidence about the true state of the world, we conjecture that a question provides evidence about the *speaker's goals*.

Rather than defining the meaning of a question in terms of its possible answers, as in traditional

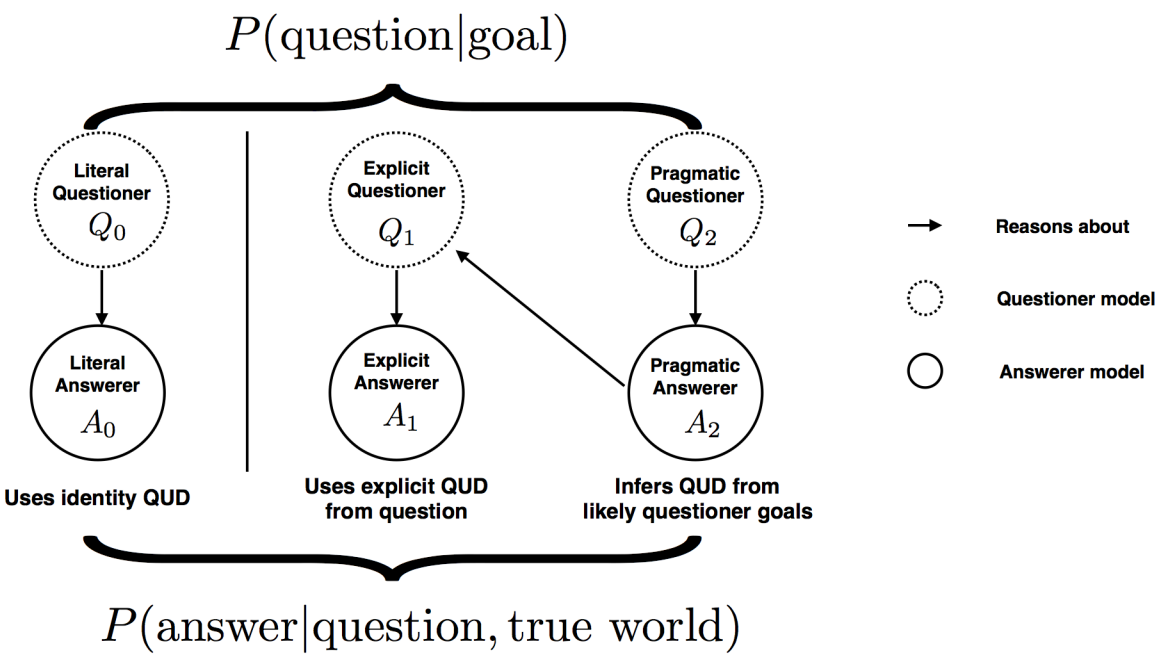

Figure 1. Schematic of the questioner and answerer models we consider, showing their recursive relationships. Note that there are two possible "base cases" where the recursion can terminate: in a *literal* answerer that ignores the question and simply attempts to be informative with respect to the world, and in an *explicit* answerer that uses the semantics of the questions.

linguistic theories, we treat it as a *signal* that the listener can use to update their beliefs about the speaker's goals. When it is the listener's turn to speak, they can then generate statements that address the goal-relevant aspects of the world. In the next section, we review the basic RSA framework and explore several different proposals for how agents interpret and respond to questions within this framework.

Preliminaries

Our mathematical formalization of dialogue begins with four primitive sets that agents take to be in common ground. For concreteness, we use the scenario from Clark (1979) as an example throughout:

$\mathcal{W}$: a set of true states of the world, such as the true price of a bottle of whiskey {\$1, \$2, ..., \$10}. In principle, this may also include other features of the world, such as the current weather, the location of nearby cafés, or the kinds of payment that the store accepts.

$\mathcal{G}$: a set of possible underlying questioner goals, such as learning the actual price of the whiskey or learning whether or not one can afford to buy the whiskey at all. We formalize the key notion of an informational goal $g \in \mathcal{G}$ as a projection function $g : \mathcal{W} \rightarrow \widehat{\mathcal{W}}$ that maps a world state w to a particular feature or set of features that the questioner cares about, which we denote $\widehat{w} = g(w)$; this is similar to the notion of a question-under-discussion (QUD; Roberts, 1996). For example, the customer's goal of learning the actual price of the whiskey is given by the identity function $g_{=}(w) = w$, while the goal of learning whether it is affordable is given

by

$$g_<(w) = \begin{cases} 1 & w < 5 \\ 0 & \text{otherwise} \end{cases}$$

such that all worlds where the agent can afford the whiskey are mapped onto one state, and all other worlds are mapped onto a different state.

$\mathcal{Q}$**:** a set of possible question utterances , such as "Does a fifth of Jim Beam cost more than \$5?" or alternatives like "Do you accept credit cards?" We take the literal meaning $[\![q]\!]$ of a question to be of the same type as informational goals, such that a particular question utterance corresponds to a particular goal projection[1]. For example, the literal meaning of the question "Does a fifth of Jim Beam cost more than \$5?" is a function that maps all worlds w such that $w >$ \$5 to one value and all other worlds to a different value. Note that the correspondence between questions and goals is not necessarily symmetric: while every question corresponds to a goal, we do not expect that every possible goal corresponds to a question utterance.

$\mathcal{A}$**:** a set of possible declarative utterances, such as "It costs \$7." or "Yes, it costs more than \$5." We take the literal meanings of these declarative utterances $[\![a]\!]$ to be truth-conditional: a map from world states to booleans. For instance, the utterance "A fifth costs \$8." would be true in the world where $w =$ \$8 and false otherwise.

A_0: Informative Answerer

We begin by considering the problem of how a dialogue agent ought to interpret and respond to a question utterance. The simplest baseline answerer, A_0, attempts to be informative about the state of the world without considering his partner's goals. In other words, A_0 is a rational speaker without any corresponding listener component to interpret and relevantly address its input.

Formally, A_0 hears a question utterance $q \in \mathcal{Q}$ with private knowledge about the true world state $w \in \mathcal{W}$. He then uses a soft-max decision rule to choose an utterance $a \in \mathcal{A}$ proportional to the epistemic utility of that utterance to his communication partner:

$$P_{A_0}(a|q, w) \propto \exp(\alpha U_{A_0}(a; w))$$

where α is a soft-max optimality parameter.

We define the epistemic utility of the utterance using the information-theoretic measure of surprisal: the increase in an imagined interpreter I's certainty about the true state of the world after hearing a:

$$U_{A_0}(a; w) = \ln P_I(w|a) \tag{1}$$

Critically, A_0 assumes that I uses Bayesian inference to update their beliefs about worlds, conditioning on the literal meaning of the utterance a being true:

$$P_I(w|a) \propto \delta_{[\![a]\!](w)} P(w)$$

where δ_e is the delta function returning 1 when e evaluates to true and 0 otherwise, and $P(w)$ is a prior over true states of the world.

This formulation of A_0 is equivalent to the pragmatic speaker S_1 for declarative statements in previous RSA models (Goodman & Frank, 2016). It thus serves as a useful baseline model for several reasons: for one, it chooses utterances using the same basic utility as the more sophisticated models we consider, thus serving as a "lesioned" speaker to evaluate the necessity of the listener component. At the same time, it is a more charitable baseline than a purely random speaker: if asked about a fifth of Jim Beam knowing that it costs \$8, A_0 would

[1] This is equivalent to the more common partition semantics of Groenendijk and Stokhof (1984), as can be seen by considering the pre-image of such a question projection: $q^{-1}(\hat{w})$.

prefer "It costs \$8" to "It costs more than \$5" or (falsely) "It costs \$4" simply because the former leads to an interpreter placing more probability on the true state of the world (w = \$8).

A_1: Relevantly Informative Answerer

While A_0 is informative, he is not particularly *relevant*. When asked about the weather, he is just as likely to inform his partner about what he ate for breakfast as he is to say "it's sunny." How do we incorporate relevance into the answerer's utility?

Following Kao, Wu, et al. (2014), we formalize speaker relevance by means of a projection function. Instead of trying to increase the listener's certainty about the most fine-grained true world state, a relevant speaker only cares about the listener's certainty in a coarser space in which irrelevant features have been collapsed together. That coarser space is the image $\widehat{\mathcal{W}}$ of the goal-projection g. Thus, we modify the utility from Eq. 1 to define a relevantly informative answerer A_1:

$$U_{A_1}(a; g, w) = \ln \sum_{w' \in \mathcal{W}} K_g(w, w') P_I(w'|a) \qquad (2)$$

where $K_g(w, w')$ is a similarity function (or kernel) between worlds, given by the goal. For the discrete space of goals considered in the current work, we use the delta function $K_g(w, w') = \delta_{g(w)=g(w')}$ which combines the probabilities of all worlds that map to the same value in $\widehat{\mathcal{W}}$[2].

This modified utility suggests simply using the semantics of the question q itself as the goal projection:

$$P_{A_1}(a|q, w) \propto \exp(\alpha U_{A_1}(a; q, w))$$

This implements the most common theory of answerers in dialogue: A_1 directly interprets a question as an epistemic goal, then informatively produces an utterance that reduces the questioner's uncertainty under that goal projection.

A_1 may be sufficient for many simple, fact-based questions, like "Who was the 16th president of the U.S.?" given a rich enough semantic parser to supply the semantics of q (Berant et al., 2013). Still, the dazzling display of context-sensitivity and indirectness displayed in everyday communication suggests that deeper social reasoning may be at play. To explore more sophisticated answerers that reason about context and their partner's underlying goals, we must first address the problem of how a speaker chooses between questions.

Q_i: Questioners

If the utility of uttering a declarative statement is imparting information about the true state of the world, what is the utility of uttering a question? We begin by assuming that a 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 her dialogue partner would respond in different possible states of the world. She then selects a question that results in an answer that tends to reduce her own uncertainty about goal-relevant information.

More formally, a questioner agent takes a goal $g \in \mathcal{G}$ as input and returns a distribution over question utterances $q \in \mathcal{Q}$ by solving a simple planning problem involving two components: (1) which answers are likely to come back after asking this question, and (2) how much would be learned from each of those answers. Critically, in order to solve either of these parts, the questioner must have in mind an imagined answerer who hears her question and responds appropriately: thus, we have a family of questioners Q_i, each using the corresponding answerer A_i:

$$\begin{aligned} P_{Q_i}(q|g) &\propto \exp\{\alpha U(q; g)\} \\ U(q; g) &= \mathbb{E}_{P(a|q)}\left[IG^g(a, q)\right] \\ &= \textstyle\sum_{w \in \mathcal{W}} P_{A_i}(a|q, w) P(w) IG^g(a, q) \end{aligned}$$

[2]Note that g-projection is a generic operation on any probability distribution $P(w|\cdot)$, which we denote:

$$\widehat{P}^g(w|\cdot) = \mathbb{E}_{P(w'|\cdot)}\left[K(w, w')\right]$$

The expectation over likely answers $P(a|q)$ is computed by running an imagined answerer model forward, marginalizing over the possible worlds $w \in \mathcal{W}$ they might have knowledge of. The value of an answer depends on its *information gain* $IG^g(a, q)$: the gap between the questioner's beliefs about the g-relevant state of the world before and after hearing an answer. Drawing again upon information theory, we formalize the notion of information gain using the Kullback-Leibler divergence between the g-projected prior distribution and posterior after hearing a.

$$IG^g(a, q) = D_{KL}(\widehat{P^g}(w|q, a) \,||\, \widehat{P^g}(w))$$

In order to anticipate her posterior beliefs after hearing an answer, the questioner must again make use of an imagined answerer, but now imagining herself as a listener in the future and inverting that model to infer the true world the answerer was trying to communicate:

$$P(w|q, a) \propto P_{A_i}(a|q, w)P(w)$$

Note that Q_0, who plans their question by reasoning about the likely responses of A_0, does not prefer any utterance over any other, because they will all lead to the same distribution of (informative) answers. Q_1, on the other hand, expects A_1 to directly interpret her question as a goal and respond informatively and relevantly, thus allowing for some questions to be better than others. It therefore instantiates the same expected information gain measure of usefulness as proposed by current Optimal Experiment Design (OED) models of human inquiry (Coenen et al., 2017). Critically, in contrast to Groenendijk and Stokhof (1984) and Van Rooy (2003), the questioner's behavior is not governed fully by the semantics of the question she asks, but by what she actually expects her partner to say after hearing it. They may be the same when reasoning about the simple A_1 answerer, but we next turn to a more socially perspicacious answerer where they diverge.

A_2: Goal-Sensitive Answerer

Now that we have defined the utility of asking a question, we can construct our final pragmatic answerer: A_2. The speaker component of this answerer attempts to informatively and relevantly address the questioner's goal like A_1. The listener component, however, accounts for the generative process of an asker with an underlying goal rather than taking that goal to be identical to the literal meaning of the question. That is, A_2 assumes that the question was produced by a rational agent Q_1, who is trying to address a private goal $g \in \mathcal{G}$, and then attempts to be relevantly informative with respect to his posterior over *underlying* goals $P(g|q)$ given the question:

$$U_{A_2}(a; q, w) = \sum_{g \in \mathcal{G}} P(g|q)\widehat{P^g_I}(w|\, a) \qquad (3)$$

Reasoning backwards from questions to goals is a simple Bayesian inversion of Q_1 using a prior on goals:

$$P(g|q) \propto P_{Q_1}(q|g)P(g)$$

By defining A_2 recursively in terms of Q_1, we implicitly give rise to an arbitrarily deep chain of nested reasoning where A_i infers a posterior over underlying goals by inverting Q_{i-1}, and so on. This recursion terminates in the base case A_1, which uses the literal question semantics. For our purposes, however, we only consider models up to a depth of $i = 2$, after which there are no qualitative changes in the model's behavior in the cases we examine[3]. In particular, we are interested in Q_2, which can be interpreted as an agent who rationally selects a question not simply for its literal meaning but for the signal it provides about her

[3]Recent systematic model exploration by Frank, Emilsson, Peloquin, Goodman, and Potts (2017) suggests that recursive depth trades off with the soft-max optimality parameter α. Because these two parameters are not generally identifiable, we fix the maximal level of recursion and achieve the behavior of higher levels of recursion by allowing α to vary.

underlying beliefs. This concludes our specification of the model space, giving a set of three answerers and three corresponding questioners that reason about them (see Fig. 1)[4].

Case studies in answerer sensitivity

In this section, we illustrate our *answerer* models through several classic goal-sensitivity effects reported by Clark (1979). These case studies center around three empirical observations about the factors determining answerer behavior in dialogue. First, and most obviously, answers should be sensitive to the surface form of the question utterance: "What time do you close tonight?" should elicit a different distribution of responses than "What is the price of a fifth of Jim Beam?". Second, answers should be sensitive to context: the *same* question utterance elicits different responses across contexts where different goals are more or less likely. Third, and most subtly, answers should be sensitive to the cue provided by the question utterance itself toward the questioner's underlying goals[5].

Sensitivity to question utterance

The minimal requirement for any answerer model is that it behaves differently in response to different questions. This can also be considered a special case of the minimal requirement for a dialogue agent: that their behavior at time t is in some way dependent on their partner's at $t-1$. Consider the problem faced by a liquor store cashier answering one of the phone calls reported by Clark (1979): they may be asked one of two questions: q_t ("What time do you close tonight?") and q_p ("What is the price of a fifth of Jim Beam?"). How do our answerer models respond to each of these questions?

To model this situation, we take a world $w \in \mathcal{W}$ to be a tuple containing the closing time and the price of Jim Beam at the store in question. Thus, $\mathcal{W}$ is the set of all possible tuples $w = (t, p)$, where t is a time in the (simplified) set {9pm, 10pm} and p is a price in {\$4, \$5, \$6}. Similarly, we consider a simplified set of answers $\mathcal{A}$: an answerer may either state a time a_{t_i} ("We close at 9:00.") or the price of Jim Beam a_{p_i} ("A fifth costs \$5."). These utterances evaluate to true in a particular world w if and only the stated dimension has the stated level.

Our informative answerer A_0 does *not* reply differently to the two questions. In either case, he slightly prefers to inform the questioner about the price using a_{p_i} because he reasons that a literal interpreter I, after hearing such an utterance, would be left with uncertainty over only two possible worlds (where the closing time is either 9:00 or 10:00) rather than the three possible worlds consistent with an utterance stating the true time (where the price is either \$4, \$5, or \$6).

Both A_1 and A_2, on the other hand, appropriately give different answers to the two questions, though for different reasons. A_1 gives different answers because the two questions simply have different literal meanings: q_t projects a world $w_i = (t_i, p_i)$ to its first element:

$$q_t(w_i) = t_i$$

whereas q_p projects w_i to its second element p_i. In this example, because we take the space of goals $\mathcal{G}$ to contain these same two projections, A_2 makes essentially the same inference. He reasons it's more likely, via Bayes' Rule, that a questioner agent Q_1 who chose to ask q_t would have the goal of learning the closing time than the price, and updates his beliefs accordingly.

These answerers then choose an utterance that *relevantly* addresses the goal they inferred. a_{p_i}, for instance, would leave the questioner with perfect information about p_i but gives no information at

[4] We have implemented these models in WebPPL, a probabilistic programming language (Goodman & Stuhlmüller, electronic). Reproducible experimental materials as well as code for all reported analyses and model simulations is available at `https://github.com/hawkrobe/Q_and_A`

[5] These examples are implemented and runnable at `forestdb.org/models/questions-answers.html`

all about t_i. Thus, it would be a preferred response to q_p and a dispreferred response to q_t. This shows the desired sensitivity to question.

Sensitivity to context

Next, we show how our models can provide different—sometimes over- or under-informative—answers to the same explicit question, depending on context. For this illustration, we consider the results of Experiment 4 from Clark (1979). 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"). Which of our answerer models, if any, shows a similar sensitivity to context?

Our space of possible worlds consists of possible prices for the whiskey: $\mathcal{W} = \{\$1, \ldots, \$10\}$. We consider two possible goals: $g_=$, learning the actual price of whiskey, and $g_>$, learning whether or not the agent can afford to buy the whiskey at all (i.e. whether the price is greater or less than the amount the agent has in their pocket). The set of answers $\mathcal{A}$ includes exact prices $a \in \{\$1, \ldots, \$10\}$ as well as "yes" and "no." We use an answer prior that assigns equal probability to picking the category of "yes/no" answers and of "price" answers, then uniformly draws from the possible responses within each category.

We model the context sentence as a statement which the answerer must interpret before interpreting the question; we assume that it has the effect of shifting the answerer's beliefs about likely questioner goals $P(g)$ (presumably by shifting beliefs about how much money the questioner has to spend). When the context is "I'd like to buy some whiskey," we assume that the the two goals are equally likely: $P(g_=|c) = P(g_>|c)$. When it is "I only have \$5 to spend," we assume that it is more strongly in favor the agent learning whether or not they can afford it: $P(g_>|c) = .99$.

A_0 is unable to show sensitivity to context for the same reason that he did not show sensitivity to the question utterance: he does not adjust his beliefs on the basis of what he hears. Thus, he always prefers to say the exact price (e.g. "The whiskey costs \$6") because it leaves less uncertainty about the true state of the world.

Because A_1 simply uses the literal meaning of the question, which is context-independent and not tied to the questioner's goals, he does *not* give different responses in the two contexts. In either case, he is ambivalent between giving a yes/no answer and an exact answer because they are equally successful at distinguishing whether the true price is greater than or less than \$5.

Finally, we see that A_2 is context-dependent in the same way as Clark (1979) observed. When the question is prefaced with "I only have \$5 to spend," which shifts the answerer's beliefs about goals strongly toward $g_>$ (equivalent to the literal meaning of the question), then A_2 is equally likely to give both answers for the same reason as A_1 above. However, when the question is prefaced with "I'd like to buy some whiskey," which leaves in place the answerer's uniform prior over goals, then the *exact price* answer is favored more strongly. This answer is much more informative in the case when the questioner's goal is to know the exact price, and no worse in the other case. The expected value of the non-literal answer is thus higher, and the answerer responds proportionally.

Sensitivity to relationship between goal and question utterance

The subtlest effects of answerer sensitivity go beyond the overt question utterance and context alone; given the same context, answerers may respond to one question according to its literal meaning but another using more indirect or overinformative statements. One of the key predictions of A_2 is that these effects are due to different inferences about the questioner's *underlying goals* on the basis of the questioner's choice of utterance itself.

For our final case study, we consider Experiment

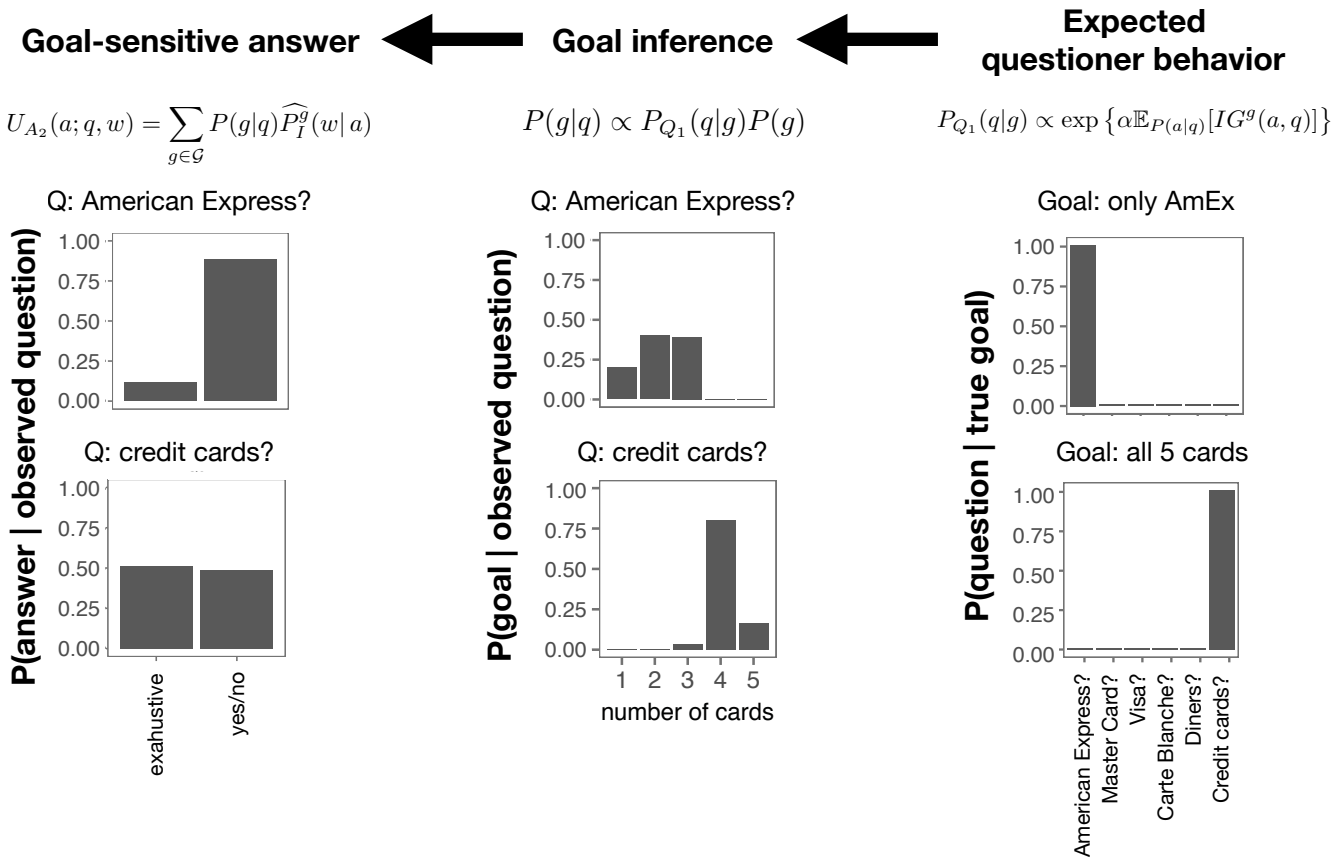

Figure 2. Illustration of how A_2 uses the questioner's utterance as a cue to their underlying goal and responds relevantly, thus producing the sensitivity reported by Clark (1979). We set the optimality parameter $\alpha = 100$ to make the effect clearer; the middle column is the marginal posterior projected to the number of cards in the goal.

5 from Clark (1979), which called restaurants and asked one of four yes/no questions about which *credit cards* the restaurant accepted. We focus on three[6]:

1. "Do you accept Master Charge cards?"
2. "Do you accept American Express cards?"
3. "Do you accept credit cards?"

Clark (1979) found that (1) and (2) were nearly always answered literally, with a 'yes' or 'no', while (3) was significantly more likely to be answered with full information. Which, if any, of our answerer models show this pattern of responses?

We formalize the scenario as follows. The set of possible worlds $\mathcal{W}$ is given by the power set $\mathcal{P}(C)$, where C = {Visa, Master Card, American Express,

[6]The fourth question was "Do you accept any kinds of credit cards?" We view Clark's results for this question as an example of M-implicature: by using a more costly utterance with the same literal meaning as question (3), the meaning becomes marked. Rational speech act models capture M-implicature by introducing *lexical uncertainty* (Bergen, Goodman, & Levy, 2012), where both speaker and listener begin their inference with uncertainty over the literal meaning of utterances. Because this is the only example of M-implicature that arises in this paper, and this is not the critical result from the experiment, we exclude it from our discussion.

Diner's Club, and Carte Blanche}, the set of five possible credit cards. The prior distribution over worlds, $P(w)$, can in principle take into account the empirical likelihoods of restaurants accepting different cards, but we treat them as uniform to make the dynamics of our simulations clearer.

The question space $\mathcal{Q}$ contains the five basic questions ("Do you accept c?" with $c \in C$ one of the five cards) in addition to "Do you accept credit cards?" For the literal semantics of the five basic questions, indexed by c, we use the function $q_c(w) = \delta_{c \in w}$, which projects worlds to a boolean corresponding to whether they accept card c or not. The latter question $q_{any}(w) = \delta_{|w|>0}$ projects to a boolean corresponding to whether any cards are accepted at all.

The answer space $\mathcal{A}$ includes lists of cards (e.g. "the cards we accept are Visa and MasterCard"), which are interpreted exhaustively, and 'yes' and 'no.' We take the denotations of these polar responses to depend on the literal meaning of the previous question: $a_{\text{yes}}(w; q) = \delta_{[\![\text{q}]\!](w)}$ and $a_{\text{no}}(w; q) = \delta_{\neg [\![\text{q}]\!](w)}$.

Finally, all goal projections $g \in \mathcal{G}$ correspond to solving a simple problem: finding out whether the restaurant accepts at least one of a set of cards of interest (i.e. "does the restaurant take any card in my wallet?"). This family of goals is parametrized by $C_q \in \mathcal{W}$, the set of cards that the questioner is actually interested in: $g_{C_q}(w) = \delta_{|C_q \cap w|>0}$.

Because A_0 does not consider the question meaning, "yes" and "no" are not well-defined: the true list of cards is the most informative answer to any question, as it provides the exact identity of the true world. A_1, on the other hand, gives either a literal yes/no answer or indirect answer in proportion to their prior probability, regardless of which question is asked. Both answers provide complete information under the question projection.

By reasoning about the questioner's underlying goals, A_2 displays the pattern that Clark (1979) observed: a higher probability of giving a literal yes/no response to "MasterCard?" and "American Express?" than to "Credit cards?". To understand *why* A_2 behaves this way, we walk through the chain of pragmatic reasoning (Fig. 2).

The key observation lies in the evidence provided by the question about the questioner's likely goal: $P(g|q) \propto P_{Q_1}(q|g)P(g)$. Consider two cases, beginning with the questioner's perspective (rightmost column of Fig. 2). First, suppose Q_1 is only interested in an American Express card (goal $g_{\{AmEx\}}$). Any answer A_1 gives to q_{AmEx} (the question "American Express?") would fully resolve her goal. "Yes" would yield certainty in accepting it and "no" in not accepting it. On the other hand, some answers to q_{any} ("Credit cards?") would not be resolving: if A_1 responded "yes", there would still be substantial uncertainty over whether the shop takes American Express in particular (for example, A_1 could informatively say "yes" if they only accepted Visa). A questioner with goal $g_{\{AmEx\}}$ is thus more likely to ask q_{AmEx}

Now, suppose instead that Q_1 has a long list of cards (say, all five: g_{all}). Then getting a yes/no for any one card would fail to address her goals: if she asked q_{AmEx} and got a "no", she is still uncertain whether another card on her list would be accepted. On the other hand, q_{any} is expected to be highly informative under A_1; if she gets a "no", she is certain that no card on her list is accepted and if she gets a "yes", she is certainty that all of them are. A questioner with goal g_{all} is thus more likely to ask q_{any}.

This asymmetry in question behavior given different goals provides A_2 information about the goal given a question: he can invert this generative model to obtain a posterior $P(g|q)$ over goals (center column of Fig. 2). By Bayes' rule, a singleton goal is more likely if the questioner asks q_{AmEx} and a longer list of cards is more likely if she asks q_{any}. The latter goal is precisely the case where an exhaustive answer is most informative (left-most column of Fig. 2), thus leading to the empirical result: answerers are more likely to give indirect, exhaustive answers when asked "Do you accept credit cards?"

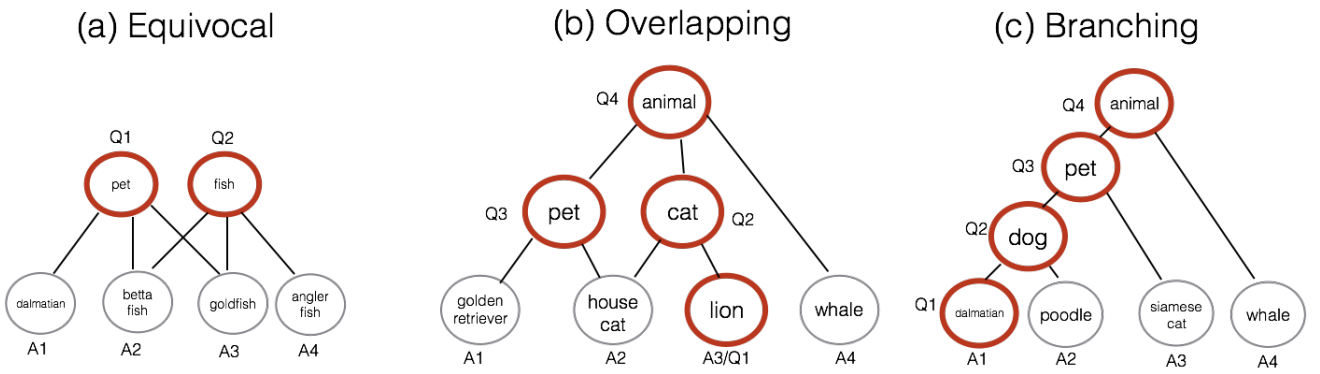

Figure 3. Example of each type of conceptual hierarchy used in our experiment. The bottom row served as the goal and answer spaces, and the labels in the questioner space highlighted in red.

Discussion

In these case studies, we illustrated three empirical answerer-sensitivity effects that serve as desiderata for how our dialogue agent should interpret and respond to questions. They should be sensitive to the question utterance, the context, and the relationship between the questioner's goal and the utterance they produce to fulfill that goal.

Our simple informative speaker A_0 was insensitive to all of these factors: it lacked a listener component that could induce any dependence on its partner's utterance. An informative answerer A_1 who also attempted to be relevant to the literal meaning of the question was sensitive to the question utterance but unable to adapt its responses to different contexts or give more or less literal answers to different questions. Finally, our most sophisticated answerer A_2, who inferred the questioner's underlying goals and attempted to be directly relevant to those goals, qualitatively captured all three effects.

The above case studies showcase the components of the modeling framework and how they explain prior experimental findings under a unified theoretical framework. However, our framework also makes novel predictions that existing data cannot address. In the remainder of the paper, we present three behavioral experiments that provide a more rigorous evaluation and extension of our framework. These experiments fill three key gaps. First, while there is extensive experimental work on questioner pragmatics is scare. directly tested whether people behave as the questioner models predict, or compared the different levels of questioner models against one another. To do so, we must collect data on which question a questioner prefers to ask given a particular goal. Second, we must provide empirical evidence for the assumptions we make about internal model choices: for instance, the priors and spaces of alternative questions, answers, or goals in a particular scenario. Finally, validate the predicted goal inference.

Experiment 1:

To satisfy these three requirements, we designed a real-time, multi-player reference game to simultaneously collect data on question-asking behavior *and* answering behavior in interaction, carefully controlling or measuring priors and utterance sets. Since Wittgenstein (1953), reference games have provided a simple but productive testbed for eliciting pragmatic behavior: one participant must choose an utterance to communicate the identity of a target object to their partner in a shared context. We extend this basic setup to a two-stage Guessing Game using natural objects and scenes as targets. On each trial, four objects are hidden such that only one player – the *answerer* – knows

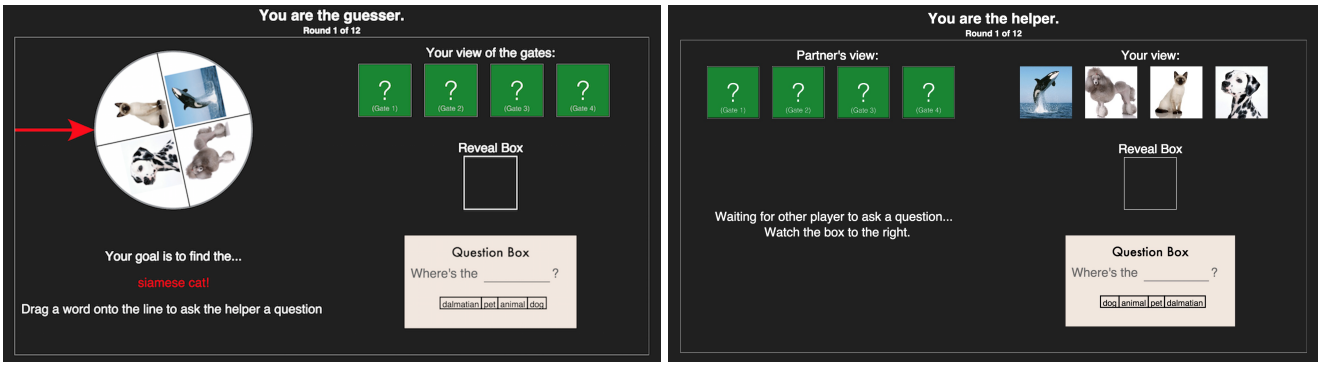

Figure 4. Experiment interfaces for the questioner (left) and answerer (right).

each object's true location. The other player – the *questioner* – is privately assigned one of the objects to find and is allowed to ask one question to query their knowledgeable partner about its location. After hearing the answerer's response, the questioner makes a guess about the location of the target object, then both players receive feedback and advance to the next trial.

The objects in context on a given trial belong to a conceptual hierarchy (see Fig. 3 for several examples), introducing a family of indirect questions. For example, a questioner can ask about the "Dalmatian," "dog," "pet," or "animal" and all would be literally true of the Dalmatian. (Brown, 1958; Graf, Degen, Hawkins, & Goodman, 2016). As discussed in detail below, our models make qualitatively and quantitatively different predictions about (1) which of these levels the questioner will choose in different contexts and (2) how answerers will respond to underspecified questions like "Where is the dog?" when there are multiple dogs in context.

Similar to the Twenty Questions task commonly used in developmental studies (Siegler, 1977; Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri, Lombrozo, Griffiths, & Xu, 2015) this game creates a knowledge asymmetry between players in order to motivate question-asking. Unlike that task, however, we withhold explicit information about the questioner's goal from the answerer. This additional asymmetry, natural to real-world social situations, provides fertile ground for pragmatic reasoning and allows us to test the necessity of goal inference in our models. Critically, both question and answer are selected from restricted sets, such that the alternative set is fixed and in common ground for all participants. In addition to allowing us to avoid ad hoc assumptions about alternative sets when comparing our models, strategic restrictions can block the most obvious, direct question and thus allow for richer pragmatic behavior. In this task, we therefore examine *which* indirect question the questioner will use, and how the answerer chooses to respond.

Methods

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 games were excluded because their participants were non-native English speakers. This left 74 unique completed games (148 participants).

Stimuli & Procedure. A set of twelve items was created by crossing four conceptual domains (animals, plants, places, and artifacts) with three concept hierarchy structures ("branching", "overlapping", and "equivocal"; see Figure 3) where our models make different predictions. Each item contains four objects (the leaves of the trees in Figure 3) which are hidden behind the four gates and may

be assigned to the questioner as goal objects, as well as four question labels drawn from different levels in the concept hierarchy (highlighted in red in Figure 3) which the questioner may use to ask about their assigned goal[7].

The procedure was designed to accommodate real-time player-to-player interaction following Hawkins (2015). All players passed a short quiz on the game instructions and were immediately redirected to the game interface where they were paired and randomly assigned to roles.

The questioner and answerer interfaces are displayed in Figure 4. At the beginning of each trial, the questioner was assigned one of the four possible goals at random (Figure 4, left). The questioner used a drag-and-drop interface to select one of four questions for the answerer, who subsequently responded with a location ("The X is behind Gate Y") by dragging the object they wanted to reveal into a box. Finally, the questioner clicked on one of the gates to make their guess, followed by feedback. Each participant provided one response for each of the twelve items, in random order.

Results

Qualitative Behavioral Results. Before conducting quantitative model comparisons, we highlight several key qualitative patterns in the behavioral data that bear on our theory. First, we note that there was generally high correlation (r = 0.91 to 0.97) in response probabilities across domains for both questioners and answerers, with the exception of the 'place' domain in answerers (r = 0.78; see Supp. Table 1). We thus collapse across domains in the following qualitative analyses, using labels from the "animal" domain for concreteness.

Next we highlight several behavioral patterns that qualitatively distinguish between our models. Suppose the questioner asks "Where is the animal?" in the branching hierarchy, where all four objects are animals. Neither A_0 nor A_1 have a preference over which of the animals to provide information about. Human answerers do. 94% of responses provided the location of the "whale," the object that isn't queried by any of the other questions, $\chi^2(3) = 210.8, p < 0.001$ (see Fig. 5A).

This strong non-uniform pattern is only consistent with A_2, which pragmatically reasons about the questioner's underlying goal given their question utterance. Because Q_1 would have been more likely to ask a different question if they had a non-"whale" goal (e.g. if looking for the Dalmatian, they would be more likely to ask "Where is the Dalmatian?"), the broad, indirect "Animal?" question is nonetheless strong evidence that their goal is to find the whale. Having inferred this goal, A_2 gives the most informative, relevant answer: the location of the whale.

Next, we turn to a subtler qualitative pattern in the "overlapping" condition (see Fig. 3), which was specifically designed to distinguish between our *questioner* models. Suppose the questioner is assigned the "house cat" as their goal object. What question do they ask? Q_0 is equally likely to ask any of the four questions, since A_0 ignores their utterance anyway. Q_1, in turn, is evenly split between the two parents in the tree – "Pet?" and "Cat?" – since A_1 would have a 50% chance of responding with the house cat's location in either case.

Q_2, however, takes into account the goal inferences A_2 would make after hearing her question. Once A_2 hears "Cat?" he will infer, using the same Gricean logic as above, that because the more specific "Lion?" is an available alternative for the questioner, she would have *directly* asked about it if it were her goal. Because she didn't, he infers that she must be interested the other cat (her true goal). "Pet?" on the other hand, gives no clue to her goals and is therefore less preferred. Thus, we have a pair of sharply distinguishable predictions: Q_0 predicts all four questions will be equally likely, Q_1 predicts that "Pet?" and "Cat?" will be equally likely, and Q_2 predicts an asymmetry

[7] 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/stimuliLabels.pdf

where "Cat?" is the preferred question.

The proportion of participants asking each possible question in this scenario, collapsed across domains, is shown in Fig. 5B. We find that the pragmatic model Q_2 makes the correct qualitative prediction – the empirical probability of asking "Cat?" ($\hat{p} = 0.64$) is significantly different from the probability of asking "Pet?" ($\hat{p} = 0.25$; bootstrapped 95% CI on difference: $[0.17, 0.58]$). This critical condition thus provides evidence for an additional layer of social reasoning: that questioners may consider the answerer's inference about likely underlying goals when selecting their question.

Quantitative model comparison. Our non-pragmatic models cannot account for key qualitative phenomena while our pragmatic models (A_2 and Q_2) make the correct predictions. Next, we show that our pragmatic models also provide better overall quantitative fits to the behavioral data. Before describing the results of our model comparison, we formalize the experimental task in our modeling framework.

Corresponding to the structure of our experiment design, we take the set of possible worlds $\mathcal{W}$ on a given trial to be the set of possible permutations of the four objects to the four locations (yielding $4! = 24$ possibilities). Because we explicitly instructed questioners to find the location of a particular object, we then take the space of goals $\mathcal{G}$ to be the set of goal projections

$$g_o(w) = \text{loc}(o, w)$$

where loc(o,w) is a simple function looking up the location of object o in world w. Similarly, the answer space $\mathcal{A}$ is the set of constructions "The o is behind gate i." which evaluate to true in worlds where o is at location i and false otherwise.

Specifying the formal semantics of the question space $\mathcal{Q}$ draws upon several concepts from linguistics. The denotation of a noun phrase containing a definite article is only defined if there is a unique, salient object that the noun picks out in the world; pragmatically, then, the use of a definite article *presupposes* the uniqueness or salience of some object in shared context (Lewis, 1979; Clark, Schreuder, & Buttrick, 1983a; Roberts, 2003). Thus we take the literal meaning of each question q to be:

$$[\![\text{ Where is the x? }]\!] = q_x(w) = \text{loc}(s_x, w)$$

where s_x is the *salient* object of category x. In contexts where there are multiple objects of category x, the question is interpreted by drawing an object from a prior saliency distribution, given the category in the noun phrase: $s_x \sim \mathcal{S}(o|x)$. This distribution may depend on typicality judgements (Rosch, 1975), world knowledge about what labels people tend to use to refer to different objects, and so on. For now, we simply take the saliency prior to be uniform over all objects that are within a category (e.g. over all leaves of the subtree beneath the given label; see Fig. 3). We assume uniform prior probability over worlds, goals, questions, and answers.

For a rigorous model comparison among our different questioner and answerer models, we conducted a Bayesian data analysis (Lee & Wagenmakers, 2014). We introduce a discrete hyperparameter $\gamma \in \{0, 1, 2\}$ determining which model to use, and jointly infer this parameter along with the continuous rationality parameter α by conditioning on our empirical data. We place uninformative priors over these parameters, $\gamma \sim \text{unif}\{0, 1, 2\}$ and $\alpha \sim \text{unif}(0, 20)$, and perform inference using enumeration over discrete bins. After integrating over values of α, the posterior $P(\gamma|\text{data})$ can be interpreted as the relative evidence for each model (Kruschke & Vanpaemel, 2015).

Results of our model comparison are shown in Fig. 5C. We can immediately rule out A_0 and Q_0, which predict a uniform distribution of responses within each condition and had a negligible posterior probability in our model comparison. Among the remaining answerer models, we found decisive evidence for A_2 relative to A_1 (BF = 4.43×10^{87}). For our questioner models, we found substantial evidence for Q_2 relative to Q_1 (BF = 4.33). Posterior predictives are shown alongside the empirical

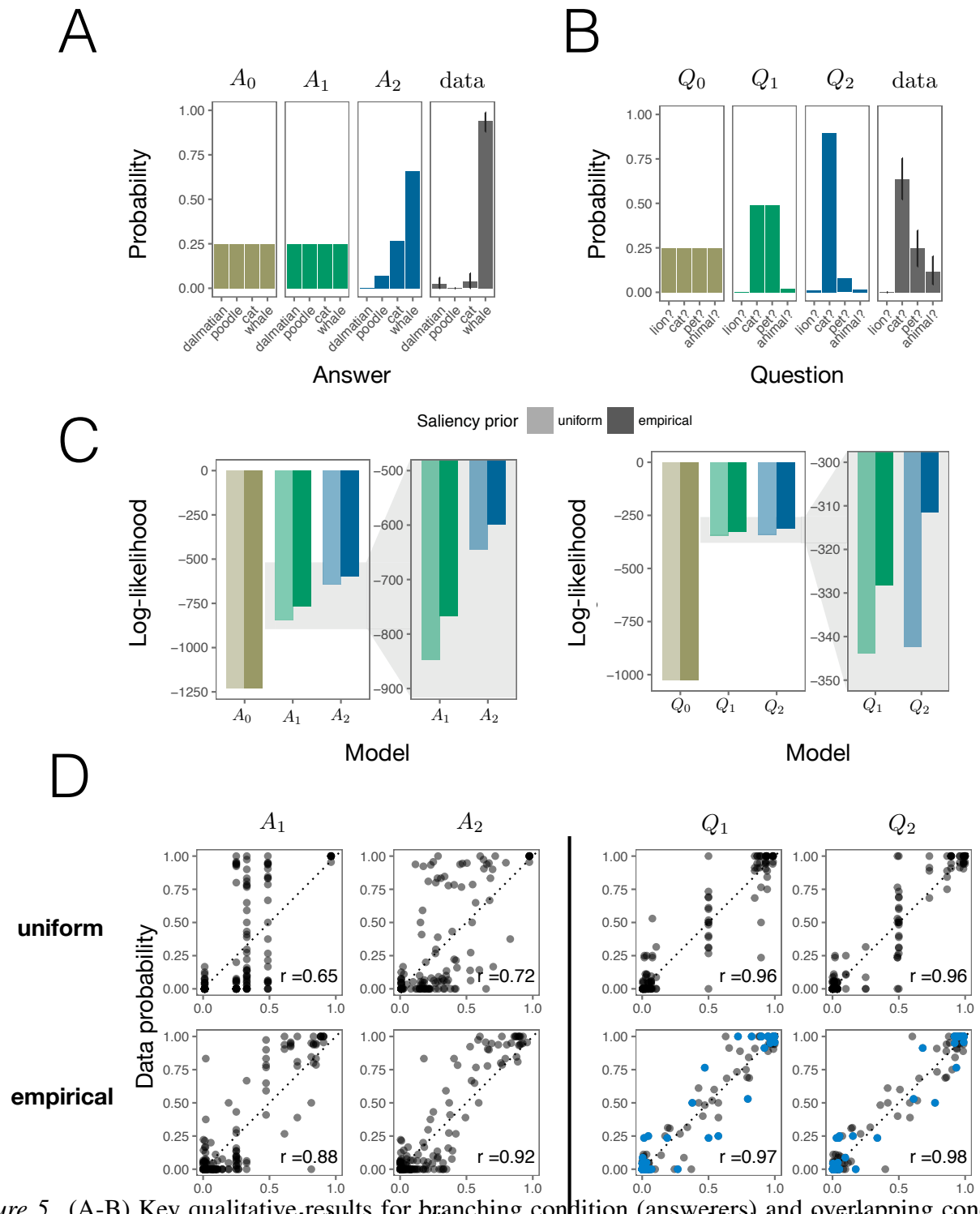

Figure 5. (A-B) Key qualitative results for branching condition (answerers) and overlapping condition (questioners). Labels shown from 'animals' domain; probabilities estimated across all domains. Error bars represent bootstrapped 95% CIs. (C) Quantitative model comparison; marginal likelihoods shown on log scale. Insert zooms in on Q_1 and Q_2. (D) Posterior predictives for each model (excluding A_0 and Q_0). Colored dots indicate overlapping condition.

data for each model in Fig. 5D.

Enriching question semantics with saliency knowledge

In three case studies, we first established that among our answerer models only A_2, which goes beyond the literal meaning of the question to infer and address underlying *goals*, could account for classic qualitative answerer-sensitivity effects. In our interactive behavioral experiment, we presented both answerer and questioner models with the additional *quantitative* challenge of accounting for trial-level variance in a reference game. We found strong converging support for A_2 relative to our other answerer models and, for the first time, tested the predictions of our questioner models. While both Q_1 and Q_2 provided good predictive fits to our question-asking data, only Q_2 could account for asymmetries in the critical *overlapping* condition and was therefore also weakly preferred in a Bayesian model comparison.

Still, Fig. 5D makes it clear that A_2 and Q_2 fail to capture graded responses across different domains. This is especially striking in the 'equivocal' condition where our models predict no preference in questioner behavior: if asking about the goldfish, which is both a "pet" and a "fish," the symmetry of the alternatives provides no pragmatic reason to favor one label over the other. Instead, we found a high degree of variability across domains. For example, when questioners were trying to find a pet goldfish, 85% of participants preferred the label 'fish' to the label 'pet,' even though both are technically true. 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.

We hypothesize that this is a typicality or saliency effect in the interpretation of the definite article (Strawson, 1950; Clark, Schreuder, & Buttrick, 1983b). A Dalmatian is a far more typical and salient member of the 'pet' category than a goldfish (Rosch, 1975), hence a questioner considering this knowledge as common ground would know that asking "Where is the pet?" would tend to elicit information about the Dalmatian, not the goldfish. If they were looking for information about the goldfish, they would then prefer to ask about the 'fish' category, where a goldfish is a highly typical example.

Failing to account for typicality could also potentially invalidate the conclusions of our model comparison. For example, we expect a house cat to be a much more typical exemplar of the category 'cat' than a lion. Hence, participants may not be choosing the house cat in this item for pragmatic reasons; they may just be selecting the more salient member of the category. If this were the case, including saliency knowledge should improve the predictions of A_1 and Q_1, thus 'explaining away' any benefit of deeper pragmatics.

We address these issues by collecting independent estimates of participants' saliency knowledge in interpreting definite articles. We then recompute our model's predictions using empirical saliency priors, in place of our previous assumption of uniform saliency within a category. We conduct a new Bayesian data analysis using these new predictions to test (1) whether our models more appropriately capture item-level variability and (2) whether the pragmatic models still provides a better fit after controlling for typicality knowledge.

Participants

We recruited 192 participants from Amazon's Mechanical Turk to participate a language-interpretation task. 13 participants were excluded after reporting a non-English native language, and 15 more were excluded due to self-reported confusion over the instructions. This left data for 164 participants.

Stimuli & Procedure

For each trial, participants were presented with a set of four pictures and asked to "Click *the X*," with *X* being some label that might refer to multiple pictures. Each participant provided a response

for twelve trials, corresponding to the twelve items from our experiment above (three hierarchy structures crossed with four domains). The four pictures corresponded to the four goal objects of the given item; we used the same pictures that were displayed on the questioner's goal wheel and the answerer's 'gates.' The instruction "Click the *X*" for a given set of pictures was formed by sampling one of the four category nouns that could be used to form a question (the highlighted nodes in Fig. 3). In this way, no participant saw the same set of objects more than once (although some images occurred in multiple sets). The order of trials and positions of pictures within trials were randomized.

Results

For each label x, we computed the proportion of participants that clicked on each of the objects o, forming an empirical saliency distribution $s_x \sim \mathcal{S}(o|x)$. To test the null hypothesis that participants use a uniform prior over within-category objects when interpreting the definite article, we performed χ^2 goodness of fit tests on the response distribution for each label: a total of 32 tests. We found that 69% of these tests rejected the null hypothesis of uniformity at a (Bonferroni-corrected) significance level of $\alpha = 0.05/32 = 0.002$ indicating that our previous assumption of uniformity was incorrect in general.

How does correcting this assumption affect our quantitative model predictions? The empirical saliency distributions we collected can be directly substituted in for $\mathcal{S}(o|x)$ when defining the literal semantics of "Where is *the*...?" questions. To test whether a model using this empirically-measured non-uniform distribution provides a better fit to the data, we add an additional parameter $\beta \sim \text{Unif}[0, 1]$ which interpolates between our original uniform salience prior ($\beta = 0$) and the empirical salience distribution we measured ($\beta = 1$). In other words, we assume participants use some mixture between the pure uniform and empirical saliency and infer the mixture weight.

First, we compute Bayes Factors marginalizing over both α and β to test whether there is still support for A_2 and Q_2 after providing the flexibility to use empirical saliencies. We again found overwhelming evidence for A_2 over A_1 ($\text{BF} \approx 6 \times 10^{72}$) and strong evidence for Q_2 over Q_1 ($\text{BF} \approx 2 \times 10^{7}$).

Next, having established support for A_2 and Q_2, we test the extent to which the introduction of empirical saliencies improved our predictions. Again using Bayes Factors, which penalize our empirical saliency model for its additional flexibility, we found strong evidence against the nested $\beta = 0$ model for both the answerer ($\text{BF} = 1.12 \times 10^{20}$) and questioner ($2.44 \times 10^{13}$), suggesting that participants interpret questions using non-uniform saliencies (see Fig. 5C). Finally, the posterior predictives for A_2 and Q_2 after including empirical saliencies explain a substantial proportion of the item-wise variance: $R^2 = .92$ and $R^2 = .98$, respectively (see Fig. 5D).

In this section, we revisited an assumption of the question semantics we used to model our guessing game experiment: that participants interpret the definite article using a uniform saliency prior. Using more realistic saliency priors elicited from a separate interpretation task, we found that this assumption was unjustified. Questioner and answerer models incorporating these more realistic priors in their meaning function provide a substantially better fit to our data; furthermore, after accounting for this potential confound, we still found strong evidence for A_2 and Q_2 over simpler models.

General discussion

Asking questions is one of our most efficient and reliable means of learning about the world. Yet we do not often pose these questions to an impartial oracle; we ask other agents, in dialogue. In this paper, we jointly consider the decision problem faced by each agent. Questioners must plan over the possible answers they are likely to receive from their partner and seek to maximize the future reduction in their own uncertainty. Answerers, in turn, seek to informatively and relevantly address

the underlying questioner goals most likely to have generated the question.

We provided several lines of evidence supporting this theory—formalized in our A_2 and Q_2 models—over purely informative answerers and asocial oracle questioners. In a series of answerer simulations, we showed that only A_2 could qualitatively account for classic answerer-sensitivity effects from Clark (1979): answers depend critically upon the question utterance, the context in which a question is asked, and the relationship between the questioner's underlying goal and their question. We then found strong evidence for A_2 and Q_2 using question and answer data jointly collected from a real-time, multi-player dialogue experiment, which was further bolstered by enriching our literal question semantics with measured saliency priors.

A major formal advance of the models considered here is synthesizing current approaches to optimal question asking and answering from AI and linguistics with the recursive social reasoning captured in the Rational Speech Act framework. In addition to introducing pragmatics into the former approaches, this moves RSA models beyond production or interpretation of single utterances (in context) to consider the dynamics of simple dialogs where both listening and speaking—linguistic input and output—must be integrated within the same agent.

Our formalization also has implications for the treatment of question semantics in linguistics. First, while the projection-based question semantics we use in our interpreter function is equivalent to classic partition-based proposals (e.g. Groenendijk & Stokhof, 1984), our *answerer* agent chooses from an independent set of declarative utterances $\mathcal{A}$, not from any set of cells induced by the question. This dissociation of the space of questions and space of answers is what allows for indirect, pragmatic responses. Second, our direct identification of question meanings with goals or QUDs raises the interesting possibility that the set of vocalizable question utterances is effectively a *subset* of the possible goals one might have: there may not exist an easy-to-express question utterance for every underlying goal an agent may need to address, thus the need for pragmatics.

Our model also complements recent work enriching active learning paradigms to allow rich natural-language questions as queries (Rothe, Lake, & Gureckis, 2016; Cohen & Lake, 2016). The expected information gain (EIG) measure used to choose new queries in these studies is closely related to our questioner formulation. Our most critical innovation to this line of work is embedding active learning in a social, communicative context (see also Shafto, Goodman, & Frank, 2012; Shafto, Goodman, & Griffiths, 2014). Instead of computing expected information gain across *all possible* answers, our questioner models take the expected value over a cooperative *dialogue partner* who seeks to be informative and relevant, often going beyond the literal question meaning. This allows questioners to effectively ask indirect questions and is especially important in real-world contexts where there is legitimate uncertainty over the questioner's goals; in a twenty-questions game, the questioner always has the goal of guessing the target object, but in a classroom context, the goal motivating a student's question may not be as obvious.

How do our assumptions scale up to less constrained interactions? We suspect that the simple, discrete sets of QUDs used in our models above will need to be generalized to continuous spaces and broader mixtures of discourse topics (Blei, Ng, & Jordan, 2003; Griffiths, Steyvers, & Tenenbaum, 2007). The computational challenges associated with such complex QUD spaces may contribute to the difficulties young children face in answering broad, sentence-focus questions like "what's happening?" (Salomo, Lieven, & Tomasello, 2013). To generalize beyond the hand-picked sets of alternative questions and answers used in our models above, we may need to consider conditional frequencies from past interactions or simple deletions and edits from the given question (e.g. Gibson,

Bergen, & Piantadosi, 2013). To generalize to dialogues lasting longer than a single exchange, we must specify the way in which the contributions of questioner and answerer affect the *context* in which later utterances operate, and the longer-term planning needed for agents to reason effectively about such interactions.

Humans are experts at inferring the intentions of other agents from their actions (Tomasello, Carpenter, Call, Behne, & Moll, 2005; Baker, Saxe, & Tenenbaum, 2009). Given simple motion cues, for example, we are able to reliably discern high-level goals such as chasing, fighting, courting, or playing (Barrett, Todd, Miller, & Blythe, 2005; Heider & Simmel, 1944). A long tradition in psycholinguistics has 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 recursive utilities of questioner and answerer agents and the computational principles by which agents infer each other's intentions from verbal behavior, our theoretical framework provides a foundation for re-situating dialogue in its social context and capturing the full richness of dialogue behavior.

Acknowledgments

We thank Leon Bergen, Judith Degen, Arianna Yuan, MH Tessler, and Judith Fan for thoughtful conversations and comments. This work was supported by ONR grants N00014-13-1-0788 and N00014-13-10341. RXDH was supported by the Stanford Graduate Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

References

Allen, J. F., & Perrault, C. R. (1980). Analyzing intention in utterances. *Artificial intelligence*, *15*(3), 143–178.

Aloni, M. (2005). A formal treatment of the pragmatics of questions and attitudes. *Linguistics and Philosophy*, *28*(5), 505–539.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*(4), 313–331.

Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In *Emnlp* (pp. 1533–1544).

Bergen, L., Goodman, N. D., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the 34th annual conference of the cognitive science society.*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Boër, S. E., & Lycan, W. G. (1975). Knowing who. *Philosophical Studies*, *28*(5), 299–344.

Brown, R. (1958). How shall a thing be called? *Psychological review*, *65*(1), 14.

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, *7*(2), 213–233.

Chouinard, M. M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, i–129.

Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive psychology*, *11*(4), 430–477.

Clark, H. H., & Schober, M. F. (1992). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about ques-*

tions: Inquiries into the cognitive bases of surveys (pp. 15–48). New York, NY: Russell Sage Foundation.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983a). Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, *22*(2), 245–258.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983b). Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, *22*(2), 245–258.

Coenen, A., Nelson, J. D., & Gureckis, T. (2017). Asking the right questions about human inquiry.

Cohen, A., & Lake, B. M. (2016). Searching large hypothesis spaces by asking questions. In *Proceedings of the 38th annual conference of the cognitive science society.*

Deits, R., Tellex, S., Thaker, P., Simeonov, D., Kollar, T., & Roy, N. (2013). Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, *2*(2), 58–79.

de Marneffe, M.-C., Grimm, S., & Potts, C. (2009). Not a simple yes or no: Uncertainty in indirect answers. In *Proceedings of the sigdial 2009 conference: The 10th annual meeting of the special interest group on discourse and dialogue* (pp. 136–143).

Der Henst, V., Carles, L., & Sperber, D. (2002). Truthfulness and relevance in telling the time. *Mind & Language*, *17*(5), 457–466.

Fong, T., Thorpe, C., & Baur, C. (2003). Robot, asker of questions. *Robotics and Autonomous systems*, *42*(3), 235–243.

Frank, M. C., Emilsson, A. G., Peloquin, B., Goodman, N. D., & Potts, C. (2017). Rational speech act models of pragmatic reasoning in reference games. *Retrieved from osf. io/f9y6b*.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gerbrandy, J. (2000). Identity in epistemic semantics. In P. Blackburn & J. Seligman (Eds.), *Logic, language and computation, vol. iii* (pp. 147–159). Stanford, CA: CSLI.

Gibbs Jr, R. W., & Bryant, G. A. (2008). Striving for optimal relevance when answering questions. *Cognition*, *106*(1), 345–369.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Ginzburg, J. (1995). Resolving questions, i. *Linguistics and Philosophy*, *18*(5), 459–527.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818 - 829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Goodman, N. D., & Stuhlmüller, A. (electronic). *The design and implementation of probabilistic programming languages.* Retrieved 2015/1/16, from `http://dippl.org`

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American educational research journal*, *31*(1), 104–137.

Graf, C., Degen, J., Hawkins, R. X., & Goodman, N. D. (2016). Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *Proceedings of the 38th annual conference of the Cognitive Science Society.*

Green, N., & Carberry, S. (1994). A hybrid reasoning model for indirect answers. In *Proceedings of the 32nd annual meeting on association for computational linguistics* (pp. 58–65).

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

Groenendijk, J., & Stokhof, M. (1984). On the se-

mantics of questions and the pragmatics of answers. *Varieties of formal semantics*, *3*, 143–170.

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966-976.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 243–259.

Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the thirty-sixth annual conference of the Cognitive Science Society.*

Kao, J. T., & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the Cognitive Science Society.*

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Kruschke, J. K., & Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. *The Oxford Handbook of Computational and Mathematical Psychology*, 279.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 1-36.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Legare, C. H., Mills, C. M., Souza, A. L., Plummer, L. E., & Yasskin, R. (2013). The use of questions as problem-solving strategies during early childhood. *Journal of experimental child psychology*, *114*(1), 63–76.

Lehnert, W. (1977). Human and computational question answering. *Cognitive Science*, *1*(1), 47–73.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of philosophical logic*, *8*(1), 339–359.

Mollá, D., & Vicedo, J. L. (2007). Question answering in restricted domains: An overview. *Computational Linguistics*, *33*(1), 41–61.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological review*, *116*(3), 499.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.

Potts, C. (2012). Goal-driven answers in the cards dialogue corpus. In *Proceedings of the 30th west coast conference on formal linguistics* (pp. 1–20).

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, 91–136.

Roberts, C. (2003). Uniqueness in definite noun phrases. *Linguistics and philosophy*, *26*(3), 287–350.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192-233.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2016). Asking and evaluating natural language questions. In *Proceedings of the 38th annual conference of the cognitive science society.*

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2015). Children search for information as efficiently as adults, but seek additional confirmatory evidence. In *Cogsci.*

Salomo, D., Lieven, E., & Tomasello, M. (2013). Children's ability to answer different types of questions. *Journal of child language*, *40*(02), 469–491.

Schulz, K., & Van Rooij, R. (2006). Pragmatic

meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy*, *29*(2), 205–250.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.

Siegler, R. S. (1977). The twenty questions game as a form of problem solving. *Child Development*, 395–403.

Simmons, R. F. (1965). Answering english questions by computer: a survey. *Communications of the ACM*, *8*(1), 53–70.

Strawson, P. F. (1950). On referring. *Mind*, *59*(235), 320–344.

Tessler, M. H., & Goodman, N. D. (2016). A pragmatic theory of generic language. *arXiv preprint arXiv:1608.02926*.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(05), 675–691.

Van Rooy, R. (2003). Questioning to resolve decision problems. *Linguistics and Philosophy*, *26*(6), 727–763.

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge University Press.

Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell.

	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
Appendix: Inter-domain correlations on corresponding response rates