

# Learning about Others: Pragmatic Social Inference through Ambiguity Resolution

Asya Achimova                      Gregory Scontras  
asya.achimova@uni-tuebingen.de      g.scontras@uci.edu

Christian Stegemann-Philipps  
christian.stegemann@uni-tuebingen.de

Johannes Lohmann  
johannes.lohmann@uni-tuebingen.de

Martin V. Butz  
martin.butz@uni-tuebingen.de

August 6, 2020

## Abstract

We investigated whether ambiguity resolution may yield socially-relevant benefits, revealing parts of the privileged ground of the interpreter. In particular, we asked if speakers can (i) use response observations to infer unknown preferences of a listener, and (ii) strategically choose ambiguous utterances for learning about those preferences. We ran experiments in a reference game framework and modeled the data with a pragmatic social inference Rational Speech Act model. Participants were able to infer listeners’ preferences when analyzing their choice of objects given referential ambiguity. Moreover, a significant group of speakers were able to strategically choose ambiguous over unambiguous utterances in an epistemic, event-predictive, goal-directed manner, although a different group significantly preferred unambiguous utterances. We conclude that ambiguity resolution indeed reveals aspects of the knowledge, preferences, and beliefs of conversation partners and some of us are able to strategically use ambiguous utterances to gain knowledge about these aspects.

**Keywords:** ambiguity; pragmatics; information gain; event-predictive cognition; Rational Speech Act models; social intelligence

## 1 Ambiguity in natural language

Ambiguity is ubiquitous during conversations: speakers rely on aspects of context and extra-linguistic reasoning to enrich the linguistic signal and deliver their

intended meanings. Given that ambiguity can hinder the efficient transfer of information between conversation partners, it is not surprising that linguists have treated the possibility for ambiguity as a bug in the communication system (Chomsky, 2002) and suggested that ambiguity should generally be avoided (Grice, 1975). If we look back at the study of ambiguity, we find that the strategy of ambiguity avoidance is much older than the pronouncements by modern linguists. Greek and Latin rhetoricians believed that a skillfully-written text allows for a perfectly accurate and lossless transmission of meaning to the reader or listener (Ossa-Richardson, 2019); such a text avoids ambiguities.

The attitude toward ambiguity has at times been quite different in other disciplines, in part because the term itself can refer to multiple phenomena. For linguistic research, a word is ambiguous if it can have two separate meanings even in the absence of context, simply as a linguistic sign. In that sense, the word “bat” is ambiguous between a winged mammal and a sporting implement. Ambiguity in the language system has been on the radars of philosophers starting with Aristotle (Sennet, 2016).

In organizational communication—communication that aids production—ambiguity aligns closely with underspecification: an utterance is ambiguous when it does not provide every detail about the intended meaning, leaving room for the listener to interpret it. This freedom is important in communication between managers and their employees when managers set future goals that should stimulate rather than limit the creativity (Mohr, 1983). Ambiguity allows expression of ideas that are true of the whole group of people, one such examples are company slogans or vision statements (Carmon, 2013). There the language needs to be vague enough to allow every member of the team to relate those general goals to oneself. Finally, ambiguous descriptions allow speakers to avoid conflict (Pascale & Athos, 1981): interlocutors find utterances that allow a range of interpretations and do not enforce a particular viewpoint.

In the case of referential ambiguity, an ambiguous utterance may apply to several possible referents in a scene. Here we use the term ‘referential’ in the sense of Frege (1892) that distinguishes the reference of a word—an object/property in the world—and its meaning. For example, a pronoun can be referentially ambiguous if there are multiple potential antecedents in the context. It is this latter type of ambiguity that we focus on in the paper, although the lessons we learn are likely to apply to the broader range of ambiguity phenomena.

In spite of the early advice to avoid ambiguity, more recent research has begun to take notice of the efficiency ambiguity affords us: by relying on context to fill in missing information, we can reuse lightweight bits of language rather than fully specifying the intended message (Levinson, 2000; Piantadosi, Tily, & Gibson, 2012; Wasow, 2015). Viewed in this way, ambiguity serves as a feature—not a bug—of an efficient communication system. This reasoning accords with years of psycholinguistic research documenting that speakers readily produce ambiguous utterances (see Ferreira, 2008, for an overview). Along related lines, Wasow (2015) reviews a large body of evidence and concludes that ambiguity is rarely

avoided, even in situations where its avoidance would be communicatively appropriate. This observation stands at odds with the Gricean maxim to avoid ambiguity (Grice, 1975). However, even Grice recognized a case of strategic ambiguity where it could be the intention of the speaker to communicate more than one possible interpretation afforded by an ambiguous utterance. In such cases, recognition of the ambiguity serves as the communicative purpose of the utterance. Wasow (2015), on the other hand, reviews several cases where ambiguous production serves no obvious communicative purpose.

In this work, we focus on the effects of resolving—or anticipating the resolution of—ambiguous utterances, identifying an additional benefit to ambiguous language: the *extra* information we gain from observing how listeners resolve ambiguity. We show that language users learn about each other’s private knowledge when observing how ambiguity is resolved. When utterances leave room for interpretation, listeners must draw on their opinions, beliefs, and preferences to fill in the gaps; by observing how it is that a listener fills in those gaps to resolve the ambiguity, speakers thus learn about the opinions, beliefs, and preferences of their conversation partner. Over the course of two studies, we first demonstrate that people are indeed able to infer hidden beliefs of their conversation partners on the basis of observed ambiguity phenomena; second, we show that some speakers can actively create situations of uncertainty, anticipating the epistemic value when observing the consequent referent choice. We thus show that humans can interpret behavior of other people as driven by their motives, intentions, or personal characteristics—an idea that goes back conceptually to the attribution theory (Jones & Davis, 1965; Kelley, 1967; Kelley & Stahelski, 1970).

To explain the behavior we observe in our experiments, we advance a computational cognitive model of the involved probabilistic inference processes formulated within the Bayesian Rational Speech Act modeling framework. While our model also pursues Bayesian inference, or “psychological reasoning”, we do not focus on the inference of the actor’s knowledge, that is, on *learning from others* (Shafto, Goodman, & Frank, 2012). Rather, we focus on *learning about others*, that is, learning about listeners’ preferences when observing how they resolve ambiguity. We explore interpretive choices and the potential strategic, socially epistemic usage of ambiguous utterances in anticipation of actors’ responses. Interestingly, the modeling results reveal good interpretive abilities but also strong individual differences when the task is to choose (ambiguous) utterances strategically for gaining social knowledge.

In the following, we first provide a computational background on referential ambiguity resolution (Section 2). In Section 3, we develop our computational models that are able to infer the preferences of an agent that led her to a particular choice of objects, as well as a model that predicts which utterances are most useful to create the possibility of learning about the preferences of the conversation partner. Sections 4 and 5 give the results of the behavioral experiments, as well as an evaluation of modeling performance. Section 6 concludes that participants were indeed able to use observable behavior of others to infer their prior beliefs, and hy-

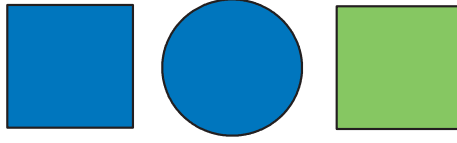


Figure 1: A simple reference game scenario from Frank and Goodman (2012). In the game, speakers are confronted with a collection of objects, which determine the current scenario  $S$ , where  $S = \{\text{solid blue square}, \text{solid blue circle}, \text{solid green square}\}$  in the depicted example. A speaker may choose a single-word utterance  $u$  to signal one of the objects  $s \in S$  to a listener. In the shown scenario, the following set of utterances is available:  $U = \{\text{“solid”}, \text{“blue”}, \text{“green”}, \text{“square”}, \text{“circle”}\}$ .

pothesizes why the ability to intentionally create epistemic situations can be found only in part of the population.

## 2 Probabilistic modeling of ambiguity resolution

To see the potential epistemic benefit of ambiguous language, take the scenario in Figure 1. Suppose a speaker produces the single-word utterance to signal one of the objects to a listener. Upon hearing “blue,” the listener faces referential ambiguity: the speaker could mean the blue square or the blue circle. Suppose further that, upon hearing “blue” in this scenario, the listener selects the blue circle. In observing this choice, the speaker learns something about the private thoughts of the listener: what made her select the blue circle instead of the blue square? Perhaps the circle is more salient to the listener, or the listener has a preference for circles, or the listener may believe that the speaker has a preference for circles; there may even be mutual agreement that circles are to be preferred when possible. Importantly, by observing how the listener resolves the ambiguity in reference, the speaker can learn something about the private thoughts of the listener.

However, accessing this added information requires the speaker to reason pragmatically about the pragmatic reasoning of the listener—a higher-order pragmatic reasoning. In order to select a referent, the listener must interpret the utterance. We follow Frank and Goodman (2012) in treating this interpretation process as active pragmatic, probabilistic reasoning: the listener interprets an utterance by reasoning about the process that generated it, namely the speaker, who selects an utterance by reasoning about how a listener would interpret it. Frank and Goodman model this recursive social reasoning between speakers and listeners, introducing the Rational Speech Act (RSA) modeling framework (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016).

## 2.1 Original RSA Formalization

Frank and Goodman’s RSA model of the reference game in Figure 1 formalizes a state space, or scenario,  $S$ , as a particular set of objects (cf. the example in Figure 1). The model unfolds computations over the corresponding utterance space  $U$ , which consists of the set of possible utterances. At the base of the reasoning process, there is a hypothetical, naïve literal listener  $L_0$ , who hears an utterance  $u \in U$  and attempts to infer the object  $s \in S$  that  $u$  is meant to reference.  $L_0$  performs this inference by conditioning on the literal semantics of  $u$ ,  $\llbracket u \rrbracket(s)$ , which returns *true* (i.e., 1) for those objects that possess the uttered feature and *false* (i.e., 0), otherwise. As a result, object choice probabilities for the literal listener can be computed by:

$$P_{L_0}(s \mid u) \propto \llbracket u \rrbracket(s), \quad (1)$$

essentially returning a uniform distribution over those objects in  $S$  that contain the uttered feature  $u$ .<sup>1</sup>

One layer up in the reasoning chain, the speaker  $S_1$  observes the scenario  $S$  and is assumed to have the intention to refer to a particular object  $s \in S$ .  $S_1$  chooses an utterance  $u$  on the basis of its expected utility for signaling  $s$  in the scenario  $S$  to  $L_0$ ,  $U_{S_1}(u; s)$ :<sup>2</sup>

$$U_{S_1}(u; s) = \log(P_{L_0}(s \mid u)). \quad (2)$$

Depending on a “greediness” factor  $\alpha$ , the speaker chooses a particular utterance  $u$  with a probability that is exponentially proportional to the utility estimate:

$$P_{S_1}(u \mid s) \propto \exp(\alpha \cdot U_{S_1}(u; s)). \quad (3)$$

At the top layer of the vanilla RSA model, the *pragmatic* listener  $L_1$  infers posteriors over  $s$  on the basis of some observed utterance  $u$ . However, unlike  $L_0$ ,  $L_1$  updates beliefs about the world by reasoning about the process that *generated*  $u$ , namely the utterance choice of speaker  $S_1$ . In other words,  $L_1$  reasons about which object  $s$  would have been most likely to lead  $S_1$  to utter  $u$  given the scenario  $S$ :

$$P_{L_1}(s \mid u) \propto P_{S_1}(u \mid s) \cdot P(s). \quad (4)$$

Frank and Goodman (2012) tested the predictions of their RSA model against behavioral data from reference games as in Figure 1. To model production behavior (that is, which utterance should be chosen to communicate a given object), the authors used the probability distributions from  $S_1$ . To model interpretation behavior (i.e., which object the speaker is trying to communicate on the basis of their utterance), the authors generated predictions from  $L_1$ . Frank and Goodman found strong correlations between model predictions and behavioral data in both cases, confirming the validity of their model of pragmatic reasoning in reference games (see also Qing & Franke, 2015, for a fuller exploration of the modeling choices).

<sup>1</sup>Note that the context  $S$  is typically not made explicit, but rather treated implicitly in the specification of the model.

<sup>2</sup>The original model in Frank and Goodman (2012) also includes a term for the utterance cost,  $C(u)$ . We ignore the term here since we assume uniform cost over all utterances.

## 2.2 Bayesian Theory of Mind

In order to interact with the environment in a highly adaptive, versatile manner, we need to anticipate how objects and agents behave in that environment (Butz, 2016). Learning about the goals, beliefs, and preferences (i.e., the priors) of other agents thus allows building predictive generative models of other agents' behavior. This process depends on the ability of humans to infer hidden states by observing behavioral choices, thereby engaging in Theory-of-Mind reasoning. Shafto et al. (2012) developed a Bayesian model of learning that formalizes the process of inferring others' knowledge about the world based on their actions and goals. They argue that efficient learning is possible if we assume that agents' actions are driven either by physical (non-social) or communicative goals, but are crucially not random. The authors show that an observer can draw stronger inferences concerning an underlying hypothesis when the agent has a communicative goal. Their model predicts that learners use knowledge of agents' goals to evaluate how knowledgeable the agents are, and, as a consequence, how much a learner can trust the agents' actions to be informative about a hypothesis.

The process of mentalizing has been modeled within the Bayesian framework in a number of papers that look at the interpretation of rational behavior of agents. Jara-Ettinger, Gweon, Schulz, and Tenenbaum (2016) review a large body of experimental work with children and adults, and propose a naive utility calculus model of so-called 'commonsense psychology'—the ability of people to infer hidden causes of behavior of others by treating them as utility maximizing agents. The ability to infer other's preferences observing their behavioral choices develops early in childhood Lucas et al. (2014), drawing from work in psychology and economics, formalize this process with a Mixed Multinomial Logit model which is driven by the assumption that in making choice agents maximize the subjective utility.

Baker, Jara-Ettinger, Saxe, and Tenenbaum (2017) develop a computational model of inference that relies on the observation of movement paths of agents to infer their beliefs, desires, and percepts. In their experiments, subjects observed an agent moving through a complex landscape looking for her favorite food-truck. According to the set-up, only two food-trucks (Mexican, Lebanese, or Korean) are allowed to be on campus on any given day. Subjects observing the movement of an agent between the present food-trucks interpret the agent's behavior to infer which food truck is her favorite, and whether she believes another food-truck is hidden in a location that is out of line of sight of the agent currently. Through comparisons with other models, the authors demonstrate that the inference crucially relies on joint reasoning about beliefs, desires, and percepts, since simpler models that rely only on a subset of these components are less accurate at predicting the human judgments.

The utility maximizing approach relies on the view of agents as being rational. Evans, Stuhlmüller, and Goodman (2016) model the inference of prior preferences when agents are not fully consistent or have restricted knowledge about the choice options. The authors propose a model that can maintain uncertainty over

the inferred beliefs and thus constitute a more realistic inference model of human preferences—an essential step for building efficient Artificial Intelligence systems that can learn about the users.

In a communicative setting, inferring hidden mental states of the conversation partners can take form of inferring priors. Thus, Degen, Tessler, and Goodman (2015) explored how listeners infer the speakers' prior over the world states and modeled the inference process within the RSA framework. If a listener is confronted with an utterance like 'Some marbles sank', she infers that the likely prior over the world states (if marbles are thrown into the water they all sink) is likely incorrect, and reason that the world must be 'wonky'. Degen et al. (2015) implement the wonkiness parameter as a lifted variable that reflects whether the listener switches to a uniform prior over the world states in case the world is wonky depending on the utterance they hear and the prior over the world states.

In a lexicon-learning paradigm, Woensdregt, Kirby, Cummins, and Smith (2016) model how Bayesian inference about the beliefs of speakers can co-develop with the process of determining the likely referents for the lexical items. The authors treat word learning as a process of inferring the intended meaning from hearing a word and observing the context in which that word was used. In a series of simulations, the model learns a lexicon and a perspective jointly. Perspective in that model is defined as salience of an object for the speaker, with salience inversely related to the distance between the speaker and the intended object. Woensdregt et al. further discuss what implications the ability to mentalize carries for reducing referential uncertainty and successful vocabulary learning.

Within the RSA framework, Yoon, Frank, Tessler, and Goodman (2018) model the comprehension of utterances describing an assessment by inferring a prior distribution over conversation goals which include not only the aim to transfer information accurately but also social goals, such as appearing nice and letting the conversation partner 'keep the face'. In work on metaphorical language use Kao, Bergen, and Goodman (2014) also consider affective goals that justify non-literal interpretation of their utterances. The authors further model the comprehension of nonliteral language as a joint inference of the meaning of an utterance and an affective state of the speaker, showing an example of a Bayesian theory of mind modeling within the RSA framework.

Our work contributes to modeling of prior inference in communicative settings within the RSA framework as well. Unlike previous papers that model the listener's inference over the priors of the pragmatic speaker  $S_1$  (Degen et al., 2015; Kao et al., 2014), we aim at modeling the inference process of a higher order pragmatic speaker upon having observed a choice of an object. Our model infers an informative distribution over preferences of the listener, rather than resorting to a uniform prior in case the listener decides the world is wonky (Degen et al., 2015). In addition to simulating the prediction of a model, we compare the model predictions to behavioral data, continuing the tradition of RSA framework and setting this paper in contrast to prior inference modeling work that relies on simulations of the interaction between two artificial agents (Woensdregt et al., 2016; Blokpoel

et al., 2019).

We further explore ambiguity use as a strategic tool that speakers may use to probe the listener to make an informative decision and reveal their preferences in a particular domain. Continuing the line of research, such as Yoon et al. (2018) and Kao et al. (2014), we consider communication goals that go beyond pure information transfer. However, unlike Hawkins, Stuhlmüller, Degen, and Goodman (2015), the epistemic goals of the speaker concern not the state of the external world but rather the mind of the listener. This aspect of our work relates to the predictive mind perspective and the active inference paradigm (Friston et al., 2015; Clark, 2016).

### 3 Our model of social inference

Our model builds on the vanilla version of RSA, modifying the listener’s state prior  $P(s)$  and enhancing the reasoning process towards a social component, yielding a *pragmatic social inference RSA* model (PSIRSA). By changing  $P(s)$  to a non-uniform distribution, we essentially model prior beliefs of which object the speaker is more likely to refer to, or—when viewed from a more self-centered perspective—which prior object feature preferences  $f$  the listener may have. For example, the listener may like blue things, such that she may be more likely to choose the blue square instead of the green one when hearing the utterance “square” in the scenario shown in Figure 1. As a result, when a pragmatic speaker produces utterance  $u$  and observes the listener’s referent choice  $s$ , the speaker may infer posteriors over possible feature preferences, attempting to explain the observed object choice in this way.

We use  $L_0$  and  $S_1$  from the vanilla model, but we now parameterize  $L_1$ ’s state prior such that it operates given a feature preference  $f$ :

$$P_{L_1}(s | u, f) \propto P_{S_1}(u | s) \cdot P(s | f). \quad (5)$$

We then model a pragmatic speaker  $S_2$ , who updates beliefs about  $L_1$ ’s preferences,  $P(f)$ .  $S_2$  observes  $L_1$ ’s choice of  $s$  given the produced utterance  $u$  and then reasons about the likely feature preference  $f$  that  $L_1$  used to make the observed choice:

$$P_{S_2}(f | u, s) \propto P_{L_1}(s | u, f) \cdot P(f). \quad (6)$$

We also model the reasoning process by which a speaker may select the best utterance to learn about the preferences of the listener, essentially striving to maximize expected information gain concerning the listener’s feature preferences. Starting with no knowledge of the listener’s preferences,  $S_2$  can be assumed to expect a uniform (i.e., flat) feature preference prior  $P(f)$ . The more the speaker’s posterior beliefs about the preferences,  $P_{S_2}(f | u, s)$ , deviate from the uniform prior, the more the speaker will have learned about the listener’s preferences. We can thus model this reasoning in light of expected information gain, which can be equated



with the attempt to maximize the KL (Kullback-Leibler) divergence between the speaker’s flat prior and the expected posterior over the listener’s feature preferences  $f$ , integrating over all hypothetically possible object choices  $s \in S$ :

$$P_{S_2}(u) \propto \sum_{s: \llbracket u \rrbracket(s)=1} P_{L_1}(s|u, f) \exp(\lambda \cdot \text{KL}(P(f) \parallel P_{S_2}(f | u, s))), \quad (7)$$

where the factor  $\lambda$  scales the importance of the KL divergence term.

We evaluate two versions of the model. `fullPSIRSA` assumes the deep reasoning process integrating the full RSA formalism. It thus assumes that feature preference inference not only considers the current object choices possible, but also differentiates the choice options further with respect to their pragmatic plausibility. For example, `fullPSIRSA` includes modeling the fact that when a speaker utters “blue” in the object situation depicted in the example shown in Figure 1 and has the intention to refer to one particular object, she is more likely to refer to the blue square than to the blue circle, because in the latter case the utterance choice “circle” would have been unambiguous and thus a better utterance choice.

Recently, it has been shown that even in the original, simpler reference games, fewer layers of reasoning often perform equally well or better than more complex RSA-based models (Sikos, Venhuizen, Drenhaus, & Crocker, 2019). Accordingly, `simplePSIRSA` removes the reasoning about alternative utterances and allows the pragmatic speaker to directly tap into the (expected) interpretation of  $L_0$ , augmenting the literal listener’s choice likelihoods with the feature-preference-dependent object prior  $P(s | f)$ :

$$P_{L_0\text{-simp}}(s | u, f) \propto \llbracket u \rrbracket(s) \cdot P(s | f). \quad (8)$$

The pragmatic speaker  $S_{1\text{-simp}}$  then reasons directly about the modified literal listener  $L_{0\text{-simp}}$ :

$$P_{S_{1\text{-simp}}}(f | u, s) \propto P_{L_{0\text{-simp}}}(s | u, f) \cdot P(f). \quad (9)$$


As a result, `simplePSIRSA` ignores any indirect pragmatic reasoning considerations about which object the speaker may refer to given an utterance and a particular object constellation. It simply assumes that all objects may be chosen that match the utterance, modifying these choice options dependent on the feature-preference-dependent object choice priors. The corresponding utterance-selection model simplifies the reasoning process accordingly:

$$P_{S_{1\text{-simp}}}(u) \propto \sum_{s: \llbracket u \rrbracket(s)=1} P_{L_0}(s|u, f) \exp(\lambda \cdot \text{KL}(P(f) \parallel P_{S_{1\text{-simp}}}(f | u, s))). \quad (10)$$


In the evaluation section below, we compare the modeling performance of `fullPSIRSA` with `simplePSIRSA`.

## 4 Experiment 1: Inferring preferences







Our first task is to check the inferences of the pragmatic speaker having observed that a listener selects some object  $s$  in response to an utterance  $u$ . Is it possible to draw inferences about the most likely preferences the listener had when making her choice? Can this inference process be modeled by PSIRSA—that is, by recursive, Bayesian inference? A sample trial is shown in Figure 2.

Progress: 

Suppose Maria wants to signal an object in the following scene to Samantha.  
Maria says "red" and Samantha chooses the outlined object:



Based on this choice, do you think Samantha has a preference for certain types of objects?

	very unlikely	very likely		very unlikely	very likely
solid things			clouds		
striped things			circles		
polka-dotted things			squares		

[Continue](#)

Figure 2: A sample trial from *Experiment 1: Inferring preferences*. Each trial portrays a speaker and a listener. The speaker produces an utterance to refer to one of the objects. The listener picks the object with the orange dotted outline. Participants were tasked with evaluating what preferences of the listener may have led her to the particular object choice, specifying their inference by adjusting the sliders for each of the features.

### 4.1 Participants

We recruited 90 participants with US IP addresses through Amazon.com’s Mechanical Turk crowdsourcing service. Participants were compensated for their participation. On the basis of a post-test demographics questionnaire, we identified 82 participants as native speakers of English; their data were included in the analyses reported below. We obtained a confirmation from all the subjects that they agree to participate in the study.

### 4.2 Design and methods

We presented participants with a series of reference game scenarios modeled after Figure 1 from Frank and Goodman (2012). Each scenario featured two people and three objects. One of the people served as the speaker, and the other served as the

listener. The speaker asks the listener to choose one of the objects, but in doing so she is allowed to mention only one of the features of the target object. Participants were told that the listener might have a preference for certain object features, and participants were tasked with inferring those preferences after observing the speaker’s utterance and listener’s object choice.

We followed Frank and Goodman (2012) in our stimuli creation. Objects were allowed to vary along three dimensions: color (blue, red, green), shape (cloud, circle, square), and pattern (solid, striped, polka-dotted). The speaker’s utterance was chosen at random from the properties of the three objects present, and the listener’s choice was chosen at random from the subset of the three objects that possessed the uttered feature. By varying the object properties, the targeted object, and the utterance, we generated a total of 2400 scenes. Speaker and listener names were chosen randomly in each trial. Participants saw the speaker’s utterance in bold (e.g., “red” in Figure 2) and the listener’s choice appeared with a dotted orange outline (e.g., the center object in Figure 2). Based on the observed choice, participants were instructed to adjust a series of six sliders to indicate how likely it is that the listener had a preference for a given feature. The sliders specified the six feature values of the two feature dimensions that were not mentioned in the speaker’s utterance (e.g., pattern and shape in Figure 2).

To compare PSIRSA’s predictions to the human data, we calculated an average value for each slider. We excluded the sliders if their corresponding feature value was not present in a scene. For example, for the trial depicted in Figure 2, we excluded the sliders for solid things and squares since none of these are present, and therefore no learning about them is possible.

To determine model correlations with the gathered data, we partitioned the data into ambiguity classes, similar to Frank and Goodman (2012). Depending on how many features competitor objects share with the chosen object, we were able to identify 48 ambiguity classes, which group the constellations that have the exact same ambiguity pattern. The ambiguity classes identified in Experiment 1 distinguish how many objects are referenced by the utterance, how the referenced objects differ in their two non-uttered features, and how the non-referenced objects differ from the referenced objects and from each other. As a result, each ambiguity class yields unique model prediction values for the individual features present (with respect to their “ambiguity role” in the particular ambiguity class) in corresponding scenarios  $S$ , effectively distinguishing all model-relevant cases. Please see the Appendix for examples of different classes.

Participants completed a series of fifteen trials. Objects and utterances were chosen as detailed above, with the constraint that ten trials were potentially informative with respect to listener preferences and five trials were uninformative with respect to listener preferences (e.g., observing that the listener chose one of three identical objects).

### 4.3 Free parameters and optimization procedure

We fit the model parameters either at the individual level or at the group level by optimizing the KL divergence between the data and the model predictions:

$$\text{KL}(P_{data}(f | u, s) || (P_{model}(f | u, s))), \quad (11)$$

where  $P_{data}(f | u, s)$  specifies a participant’s normalized slider value setting, which offers empirical estimates of the feature-preference posterior given object scene  $S$ , a particular utterance choice  $u$ , and the consequent object choice  $s$ ;  $P_{model}(f | u, s)$  specifies the corresponding model posterior, either  $P_{S_2}(f | u, s)$  in the case of fullPSIRSA or  $P_{S_{1-simp}}(f | u, s)$  in the case of simplePSIRSA. By minimizing the summed KL divergence between the empirical and model-predicted preference posteriors over all considered trials, we essentially maximize the model fit to the participants’ data. Moreover, we can use the minimized KL divergence values to calculate the  $G^2$ -statistic and perform the likelihood ratio test for nested models, since  $G^2$  values are approximately chi-square distributed (Lewandowsky & Farrell, 2011). Individual vs. global parameter fitting allows us to explore potential differences between participants. In the case of individual model parameter optimization, parameters were optimized for each individual participant separately, determining the KL divergence with respect to the participant-specific set of trials. In the case of global optimization, all trials of all participants were used to determine the summed KL divergence.

We fit three parameters for fullPSIRSA and two for simplePSIRSA. The softmax scaling factor  $\alpha$  is only relevant for fullPSIRSA; it controls how likely speaker  $S_1$  is to maximize utility when choosing utterances. The default value is typically set to  $\alpha = 1$  (i.e., no scaling).

The softness parameter  $\gamma$  regulates the strength of individual feature preferences  $f$ :

$$P(s | f) \propto \begin{cases} 1 + \gamma, & \text{if } s \text{ contains } f \\ \gamma, & \text{otherwise} \end{cases}, \quad (12)$$

controlling the choice probability of those objects  $s$  that contain feature  $f$  compared to those that do not. A value of  $\gamma = 0$  models a hard preference choice; in this case, the speaker always chooses one of the preferred objects. On the other hand, when  $\gamma \rightarrow \infty$ , the choice prior becomes uniform over all objects, thus ignoring feature preferences.

For example, in the trial shown in Figure 2, there are two objects that fit the utterance  $u = \text{“red”}$ : a red striped cloud and a red dotted circle. When  $\gamma = 1$ ,  $P(s_{\text{red striped cloud}} | f_{\text{“cloud”}}) = 2/3$ , while  $P(s_{\text{red dotted circle}} | f_{\text{“cloud”}}) = 1/3$ , yielding a soft preference for clouds. We use  $\gamma = 0$ —that is, hard preferences—as the default model value.

Finally, we allow for the possibility of noise in our human data introduced by participants not following instructions. Parameter  $\beta$  models the possibility that listeners choose objects that do not pass the semantic filter of the literal listener,

allowing for non-literal interpretations that result in choosing objects whose features do not match the received utterance  $u$ . The computation is equivalent to the softness parameter above, in this case softening the object choices of the literal listener  $L_0$  towards a uniform choice over all objects present.

Again,  $\beta = 0$  models a hard object choice—that is, full obedience to the uttered instruction  $u$ —while  $\beta \rightarrow \infty$  models a uniform object choice—that is, full ignorance of  $u$ .

## 4.4 Results

### 4.4.1 Models with global optimization

We fit the following free parameters to optimize the predictions of the models. First, the full model includes a “greediness” parameter  $\alpha$  that controls how likely it is that speakers choose the best-suited utterance to signal a particular object to a listener. This parameter is absent in the simple model since it relies on fewer layers of reasoning. The second parameter  $\gamma$  controls how soft the preferences are. Hard preferences enforce the choice of the preferred object type, while increasing softness converges towards no object preference. Similarly, the obedience parameter  $\beta$  allows subjects to choose objects that do not qualify for the utterance. As for the preference parameter  $\gamma$ , the  $\beta$  range includes hard obedience on the one side of the spectrum – for example, definitely choosing a blue object when hearing “blue” – and full ignorance of the utterance at the other extreme, choosing uniformly from all available objects.

simplePSIRSA and fullPSIRSA with softness ( $\gamma$ ) optimized globally provide nearly identically good fits to the data (Figure 3). Simple linear regression analysis was used to test whether the model values predicted the human data. simplePSIRSA yields a value of  $r^2 = 0.8607^3$  ( $F(1, 190) = 1181$ )<sup>4</sup> when only softness parameter  $\gamma$  is optimized ( $\gamma = 0.2204$  after optimization). When both parameters are optimized globally, a variance estimate of  $r^2 = 0.9788$  ( $F(1, 190) = 8823$ ) is reached ( $\gamma = 0.2210$  and  $\beta = 0.2693$  after optimization), indicating that participants indeed considered (possibly subconsciously) the option to interpret utterances non-literally. fullPSIRSA yields nearly identical values. When optimizing only the softness parameter  $\gamma$ , a value of  $r^2 = 0.8568$  ( $F(1, 190) = 1144$ ) is reached ( $\gamma = 0.2231$ ). Optimizing both,  $\alpha$  and  $\gamma$ , a value of  $r^2 = 0.8607$  ( $F(1, 190) = 1144$ ) is reached ( $\alpha = 0.1797$ ,  $\gamma = 0.2205$ ). When optimizing all three parameters, fullPSIRSA yields a value of  $r^2 = 0.9772$  ( $F(1, 190) = 8170$ ) ( $\alpha = 0.2657$ ,  $\gamma = 0.2214$ ,  $\beta = 0.0030$ ).

Overall, the results show that participants are indeed able to infer the feature preferences that lead to the choice of an object. Moreover, the higher model flexibility of fullPSIRSA—controlled via parameter  $\alpha$ —does not yield any modeling

<sup>3</sup>Here and throughout the paper we report adjusted  $r^2$  values.

<sup>4</sup>All results were significant at  $p < 0.001$  level if not stated differently in the text.

improvement, implying that an approximation of the more shallow reasoning process modeled by simplePSIRSA typically unfolded in the minds of the participants.

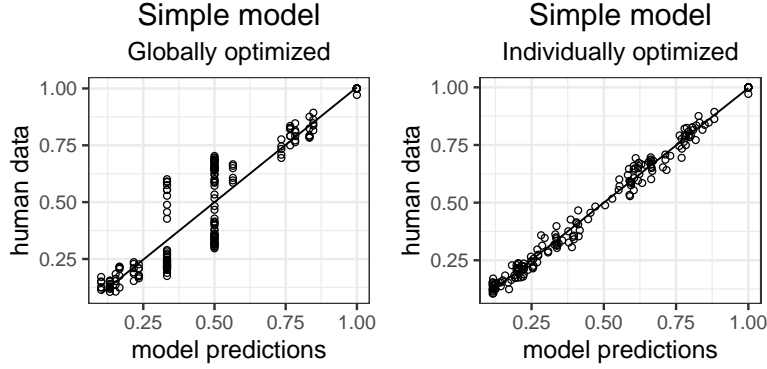


Figure 3: Human data from Experiment 1 plotted against the predictions of simplePSIRSA. Each data point indicates the slider values and model predicted feature preference posteriors for a particular ambiguity class. Left panel:  $\gamma$  *optimized globally* ( $r^2 = 0.8614$ ); right panel:  $\gamma$  and  $\beta$  *optimized individually* with leave-one-out cross-validation ( $r^2 = 0.9901$ ).

#### 4.4.2 Individually-fitted models

We now compare our two model variants further when fitting the parameters to the individual data of each participant separately. In situations when the population is potentially heterogeneous, individual level modeling in reference games improves the fit of the model despite its increased complexity (Franke & Degen, 2016). We optimized  $\alpha$  and  $\gamma$  in the light of the KL divergence between the individual participants' slider value choices and the corresponding model predictions for PSIRSA. We then again averaged the individualized model prediction values and participants' slider values with respect to the particular ambiguity classes and calculated correlations between the data and the model.

The full model optimized at the individual level for the additional parameter  $\alpha$  does not improve the fit compared to the simplified model (simplePSIRSA:  $r^2 = 0.8631$ ,  $F(1, 190) = 1205$ ; fullPSIRSA:  $r^2 = 0.8627$ ,  $F(1, 190) = 1201$ ). Seeing that both models again fit the data nearly equally well (if anything, simplePSIRSA performs slightly better), we only consider the predictions of simplePSIRSA henceforth. Note further that the individually-fitted parameters do not improve the correlation values much, if at all, when compared to the globally-fitted model.

The model fit improves considerably when we additionally fit the obedience parameter  $\beta$  at the individual level. Here the model explains a large proportion of variance in the human judgments ( $r^2 = 0.9919$ ,  $F(1, 190) = 23480$ ). The likelihood ratio test (two-tailed) revealed that a  $\gamma$ - and  $\beta$ -optimized simplePSIRSA model pro-

vides a better fit compared to a model optimized only for  $\gamma$  ( $G^2 = 237.36, df = 82, p < 0.01$ ). The more complex model contains one additional parameter  $\beta$  fitted for each subject, giving us 82 degrees of freedom. We additionally checked the generalizability of the model by performing leave-one-out cross-validation on the individual level. Figure 3 shows that the resulting cross-validated model predictions retain the strong fit ( $r^2 = 0.99, F(1, 190) = 18910$ ).

To appreciate the gains obtained by fitting model parameters, Figure 4 shows the average responses of the human participants and of the individually-, two-parameter-optimized simplePSIRSA model and the non-optimized simplePSIRSA model for the scene type of the sample trial from Figure 2. In that trial, participants saw that the middle object was chosen following the utterance “red”. There are two potential referents for this description: the red striped cloud and the red dotted circle. Since the cloud was chosen, we infer that the person who chose this object has a preference for clouds over circles, and for striped objects over dotted ones. Note that we cannot learn anything about the preference for solid things or squares in this trial because these features are not present, thus we ignore the respective slider values. Moreover, we can definitely not learn anything about color preferences because the color was uttered; thus, sliders for those features were not present. As Figure 4 shows, both humans and the models assign high slider values to clouds and striped things, and low values to circles and dotted things. Indeed, even the non-optimized model fits the qualitative pattern of the results; optimizing  $\beta$  and  $\gamma$  improves the quantitative fit.

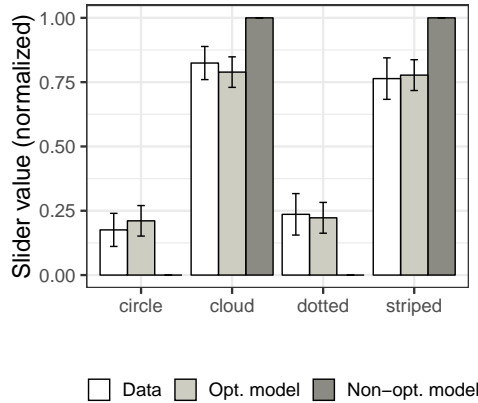


Figure 4: Human data and simplePSIRSA’s (individually-, two-parameter-optimized and non-optimized) feature preference posterior estimates for the scenario  $S$  shown in Figure 2. Error bars represent 95% confidence intervals.

We thus find strong empirical support for simplePSIRSA, implying that speakers are indeed able to use listener behavior to acquire information about their preferences. We fail to find that the fullPSIRSA model predicts the data better. This result suggests that the task in our experiments does not require full-blown prag-

matic inference about alternative utterances. The question now turns to whether speakers are able to capitalize on this reasoning when it comes to selecting utterances. In other words, are speakers aware that ambiguous language is potentially more informative and can thus use ambiguous language in a socially epistemic, strategic manner?

## 5 Experiment 2: Choosing utterances to learn about others

Our next task is to check the predictions of our strategic utterance selection model: given a set of potential referents  $S$ , are participants able to reason pragmatically about the anticipated potential epistemic utility of utterances  $u \in U$  in inferring the listener’s preferences? Figure 5 shows a sample trial, in which the speaker (“Katie” in the example) is to choose an utterance in order to learn about the listener’s preferences (“Elizabeth” in the example). While the ambiguous utterances “cloud”, “green”, and “striped” may allow inferences about color & texture, shape & texture, and color & shape, respectively, the utterances “solid”, “blue”, and “circle” leave only one response option to the listener, such that the speaker cannot learn about the listener’s preferences when observing the listener’s response (assuming the listener obeys the speaker’s order).







Progress: 

Suppose Katie wants to learn about Elizabeth's preferences in the following scenario:



Katie can choose a single utterance and then watch Elizabeth select an object.

What should Katie say?

	definitely not	definitely
"cloud"		
"solid"		
"green"		
"striped"		
"blue"		
"circle"		




Figure 5: A sample trial from *Experiment 2: Choosing utterances*.



## 5.1 Participants

We recruited 90 participants with US IP addresses through Amazon.com’s Mechanical Turk crowdsourcing service; participants in Experiment 1 were not eligible to participate in Experiment 2. Participants were compensated for their participation. On the basis of a post-test demographics questionnaire, we again identified 82 participants as native speakers of English; their data were included in the analyses. We obtained a confirmation from all the subjects that they agree to participate in the study.

## 5.2 Design and methods

Participants encountered a reference game scenario similar to Experiment 1 in which a speaker signals an object to a listener who might have a preference for certain types of objects. However, rather than observing the utterance and referent choice, participants were now tasked with helping the speaker choose an utterance that was “most likely to reveal the listener’s color, shape, or pattern preferences.”

We used the same sets of objects from Experiment 1, which could vary along three dimensions. Each trial featured a set of three objects, as in Figure 5. After observing the objects, participants adjusted sliders to indicate which single-feature utterance the speaker should choose to learn about the preferences of their listener. Potential utterances corresponded to the features of the objects present; depending on the number of unique features, participants adjusted between three and nine sliders. As with Experiment 1, we averaged the data and the respective model predictions across specific ambiguity classes, which include all scenes that yield identical utterance choice options. In this case, 14 distinct conditions can be identified, with a total of 84 slider values to set. Membership within an ambiguity class is defined by how many objects in a scene share each of the features: shape, pattern, and color. If objects share a feature, we also consider whether these objects also share other features. For example, in Figure 5, two green objects differ in shape, making the utterance *green* informative. If, on the other hand, both green objects were clouds, uttering *green* would not allow the speaker to update their beliefs about the listener’s shape preferences. In the most extreme case, when all objects share all three features, all utterances are ambiguous since multiple objects can always be picked; but no utterance allows the speaker to learn anything about the listener because the object choice is uninformative. Another extreme case is a situation where all objects are unique and do not share any features. In such a case, any utterance will only pick one object, making learning about preferences impossible unless obedience ( $\beta$ ) is not 0—that is, unless listeners have a tendency to disobey the utterance and consider objects that do not satisfy its literal interpretation.

Just like for Experiment 1, each ambiguity class yields unique model prediction values for the individual features present in the respective scenarios  $S$ , taking into account their ambiguity role. This grouping strategy effectively distinguishes all

model-relevant cases. Please see the Appendix for examples of different classes.

Participants completed a series of fifteen trials. As with Experiment 1, objects were chosen at random, with the constraint that ten trials were potentially informative with respect to the listener’s preferences (as in Figure 5) and five trials were uninformative with respect to the listener’s preferences (e.g., observing a set of three identical objects).

### 5.3 Results

We use simplePSIRSA to compute the expected most informative utterance for inferring preferences. In other words,  $P_{S_1\text{-simp}}(u)$  calculates the probability that a speaker would choose  $u$  for the purpose of inferring preferences.

To generate predictions from  $P_{S_1\text{-simp}}(u)$ , three free parameters can be identified: the preference softness  $\gamma$ , the obedience  $\beta$ , and the  $\lambda$  parameter, which factors the importance of choosing the expected most informative utterance with respect to the expected KL divergence between preference priors and expected preference posteriors (cf. equations 7 and 10). While a positive value yields the intention to maximize information gain, a negative value results in a tendency to minimize information gain, that is, a preference for no change in the posterior feature preference estimate  $P_{S_1\text{-simp}}(f | u, s)$  in comparison to the prior estimate  $P(f)$ . A value of  $\lambda = 0$  effectively ignores information gain and a resulting tendency to choose the object that was most likely referenced given the *utterance*.

We compare simplePSIRSA with non-optimized parameters and with several parameter optimizations with the performance of a uniform baseline model, which simply chooses one of the available utterances at random. Seeing that in particular ambiguity cases with particular constellations  $S$  three up to nine utterances are possible, the baseline model yields different model predictions for the available utterances in the respective ambiguity classes. As a result, the model is much better in capturing variance in the data than one would expect without this insight ( $r^2 = 0.7466$ ,  $F(1, 82) = 245.6$ ,  $p < 0.001$ ). Figure 6 compares this performance to the non-optimized simplePSIRSA, where we set the parameters to hard preference and obedience ( $\gamma = 0$ ,  $\beta = 0$ ) and the information gain factor to  $\lambda = 1$ , thus preferring to choose those utterances that are expected to yield high information gain. Surprisingly, this model captures very little variance in the human data ( $r^2 = 0.0595$ ,  $F(1, 82) = 6.253$ ,  $p < 0.05$ ).

To examine the reasons for this failure, we first performed additional global parameter optimization runs. When optimizing all simplePSIRSA parameters, the model accounts for more variance than the uniform base model ( $r^2 = 0.7991$ ,  $F(1, 82) = 331.2$ ,  $p < 0.001$ ; optimized model parameters:  $\gamma = 0.0006$ ,  $\beta = 0.2758$ ,  $\lambda = 0.3663$ ). Moreover, the nested model comparison test with three free parameters yields a  $G^2$  value of 13.6912, which indicates a more accurate model with  $p < 0.01$ . Figure 7 shows the correlation plot. The parameters indicate that the preference strength is rather high, obedience is not as strong, while the information gain intention is present. We now turn to individual parameter optimization,

suspecting that there may be fundamental differences between the individual participants.

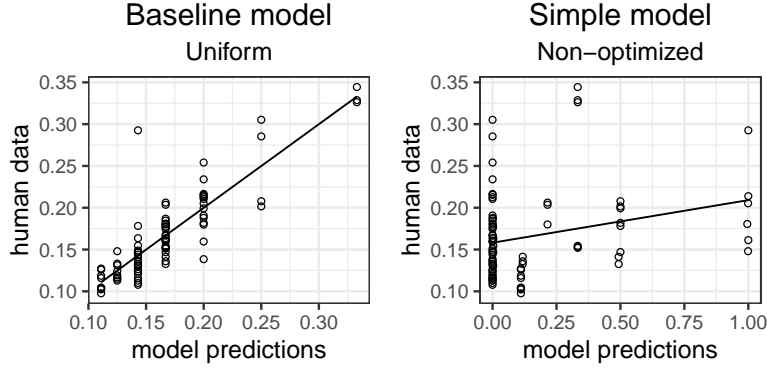


Figure 6: Average human data from Experiment 2 plotted against the predictions of the uniform baseline model and the simplePSIRSA model. Left panel: *uniform model* ( $r^2 = 0.7466$ ); right panel: *non-optimized simplePSIRSA* ( $r^2 = 0.0595$ ).

We compared three single-parameter-individually-optimized simplePSIRSA models to determine which model provides the best fit to the data. All models have similar levels of complexity, with either softness  $\gamma$ , obedience  $\beta$ , or KL-factor  $\lambda$  being optimized. The results indicate that we get the best fit by optimizing the KL-factor  $\lambda$  ( $r^2 = 0.9059$ ,  $F(1, 82) = 800.2$ ; leave-one-out cross-validated optimization  $r^2 = 0.8902$ ,  $F(1, 82) = 664.8$ , with other models capturing less variance in the data ( $\beta$ -optimized  $r^2 = 0.8015$ ,  $F(1, 82) = 336.1$ ;  $\gamma$ -optimized  $r^2 = 0.8077$ ,  $F(1, 82) = 349.6$ ). The comparison with the baseline model in terms of nested model statistics confirms that only the individual optimization of  $\lambda$  improves model performance ( $\lambda$ :  $G^2 = 268.88$ ,  $df = 82$ ,  $p < 0.001$ ;  $\gamma$ :  $G^2 = 31.38$ , n.s.;  $\beta$ :  $G^2 = 56.29$ , n.s.). Two- and three-parameter individual optimizations did not yield any significant model improvements when compared to the individually  $\lambda$ -optimized model (best improvement when optimizing  $\gamma$  in addition to  $\lambda$ :  $G^2 = 24.72$ ,  $df = 82$ , n.s.). Figure 7 shows the resulting correlation plot for  $\lambda$ -individually optimized model.

Unlike for Experiment 1, where even the non-optimized models provided a good linear fit to the data, individual optimization produces a large effect on the model predictions in Experiment 2. Figure 8 compares individually-optimized vs. non-optimized model predictions against the human behavior for the sample trial in Figure 5. We see that the non-optimized model strongly favors ambiguous utterances: in a situation with a striped green circle, a blue striped cloud, and a solid green cloud, uttering things like *cloud*, *striped*, or *green* (i.e., the utterances that point to more than one object in the scene) could let the speaker learn something about the listener’s preferences. However, Figure 8 shows that human behavior deviates quite strongly from the non-optimized, ambiguity-selecting baseline; once we optimize  $\lambda$ , we are able to capture human behavior in the task.

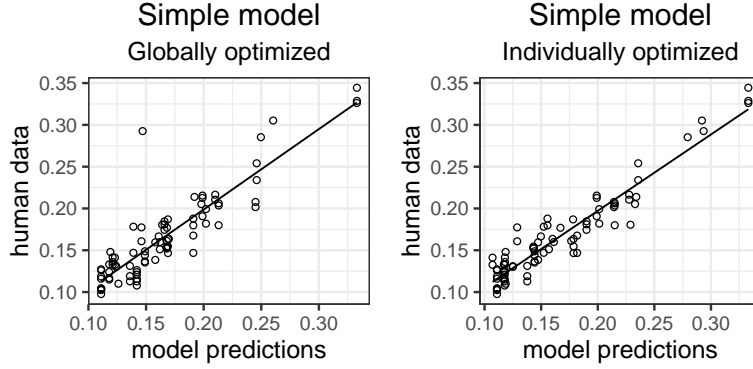


Figure 7: Average human data from Experiment 2 plotted against the predictions of optimized simplePSIRSA models. Left panel: *globally optimized 3 parameter model* ( $r^2 = 0.7466$ ; right panel: *individual KL-factor  $\lambda$ -optimized model* ( $r^2 = 0.9059$ ).

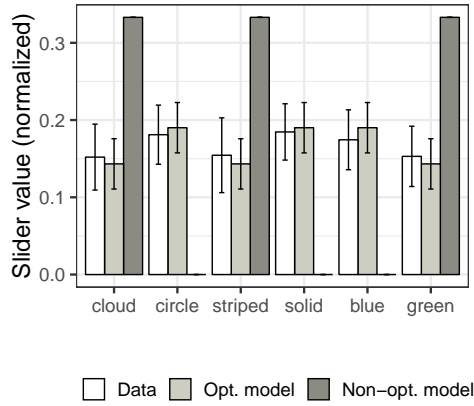


Figure 8: Simple Social Inference model predictions and human data for one of the classes of stimuli *Experiment 2: Picking utterances*. The optimized version of the model is optimized for the KL-factor  $\lambda$ . Error bars represent 95% confidence intervals.

When examining the individually optimized model values in further detail, we noticed three groups of participants. The first one may be termed a “lazy worker” group or “unpredictable” behaving group: for 28 participants, the KL divergence values of the  $\lambda$ -optimized simplePSIRSA model failed to reach the performance of the baseline model, essentially failing to identify any model-corresponding regularity in the data that goes beyond random utterance choice behavior. The second group of 33 participants yielded more negative values (i.e.,  $-7.11 < \lambda < -0.014$ ,  $\bar{\lambda} = -0.823$ ), indicating that a significant number of participants preferred to systematically choose unambiguous utterances ( $G^2 = 180, 17$ ,  $df = 33$ ,  $p < 0.001$ ).

The third group of 21 participants yielded positive values (i.e.,  $.0187 < \lambda < .537$ ,  $\bar{\lambda} = -0.124$ ), indicating that these participants indeed preferred the more ambiguous utterances in a strategic manner ( $G^2 = 102.16$ ,  $df = 21$ ,  $p < 0.001$ ).

Further experiments with highly similar setups confirmed this trend. In particular, we ran two additional, complementary studies with a blocked design where participants first completed preference-inferences trials as in Experiment 1 and then utterance-selection trials as in Experiment 2. In the first complementary study with 10 trials (135 participants, data from 123 native speakers of English included in the analysis, 12 non-native speakers excluded), the identical analysis yielded 42% participants that preferred ambiguous over unambiguous utterances (37% unpredictable participants; 21% preferred unambiguous utterances). In the second complementary study with 54 participants (2 participants excluded as non-native speakers), which contained 30 trials in total and had slightly more general instructions, as many as 64% of the participants systematically preferred ambiguous over unambiguous utterances (21% unpredictable workers; 15% preferred unambiguous utterances).

## 6 Discussion

We have found strong support that we can indeed learn about others when observing their interpretation of ambiguous utterances. The results of Experiment 1 demonstrate that naïve speakers are able to reason pragmatically about *why* listeners may take the actions they do. The success of our computational model PSIRSA in predicting the observed behavior offers an articulated hypothesis about *how* this reasoning proceeds: when speakers are aware of the ambiguity in their utterances, observing how listeners resolve that ambiguity provides clues about the preferences listeners use when doing so. The results of Experiment 2 demonstrate that at least some speakers are able to capitalize on this reasoning to strategically select ambiguous utterances that are expected to improve their understanding of the preferences of their listeners.

Currently, we are transferring the experimental setup to more naturalistic interaction scenarios. Even in these cases, though, it appears that we still find participants who consistently prefer to choose unambiguous utterances. Two explanations may be warranted and need to be investigated further. First, it may be the case that these participants think overly egocentrically, thus having the intention to signal their own preferences rather than to give options to the listener. Second, it may simply be the case that these participants do not have access to the required deeper reasoning process, and thus prefer to give instructions with predictable outcomes.

Nonetheless, taken together, the results of our experiments and the success of PSIRSA in modeling these results indicate that humans are aware of the fact that by observing responses to ambiguous utterances, information about the listener’s prior preferences can be inferred—that is, they are able to learn about the hidden model states of others, including preferences but probably also other aspects of

beliefs.

It should also be noted that ambiguous utterances used in this way are closely related to questions, which may ask directly about considered preferences. Ambiguous utterances provide a ready but more subtle, indirect alternative to asking directly. In normal conversations, a speaker might favor the indirect route, given considerations of politeness and possibly also in an effort to keep the conversation open. With ambiguous language, the conversation partner can choose to disambiguate the ambiguous utterance or, alternatively, choose to continue in a different direction or even change topic.

We note that the analyzed preference prior, viewed from a broader perspective, can be closely related to a part of the event-predictive mind of the listener and the speaker (Butz, 2016; Butz & Kutter, 2017). When interpreting an utterance—in our case, opening up a set of referent choices—the listener’s mind infers the current choices and integrates them with her preference priors, implicitly anticipating possible choice consequences. Moreover, the expected information gain term—computing the utterance choice of the speaker—can be equated with the computation of socially-motivated active inference (Butz, 2017; Friston et al., 2015). It causes the model to strive for an anticipated epistemic value that quantifies the expected information gain about the preferences of the listener—that is, expecting a form of social information gain.

More generally, predictive states of mind about others do not only include considerations of preferences, but may also concern all imaginable knowledge, opinions, beliefs, and current trains of thought of the listener. Moreover, during a conversation, the involved “social” priors will dynamically develop depending on the internal predictive models and the generated utterances, actions, and responses of the speaker and listener. The priors dynamically depend on the privileged grounds of the conversational partners, and also on the common ground in which the conversation unfolds. In that sense, ambiguous utterances and resolutions thereof are one device for projecting parts of each other’s privileged grounds into the common ground.

## **Funding**

This project has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project number 198647426.

## **Data availability**

Data supporting the findings of this study are available from the corresponding author upon request.

## References

- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4), 1–10.
- Blokpoel, M., Dingemanse, M., Woensdregt, M., Kachergis, G., Bögels, S., Toni, I., & van Rooij, I. (2019, Nov). *Pragmatic communicators can overcome asymmetry by exploiting ambiguity*. (10.31219/osf.io/q56xs)
- Butz, M. V. (2016). Towards a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology*, 7(925). doi: 10.3389/fpsyg.2016.00925
- Butz, M. V. (2017). Which structures are out there? learning predictive compositional concepts based on social sensorimotor explorations. In T. K. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573093
- Butz, M. V., & Kutter, E. F. (2017). *How the mind comes into being: Introducing cognitive science from a functional and computational perspective*. Oxford, UK: Oxford University Press.
- Carmon, A. F. (2013). Is it necessary to be clear? an examination of strategic ambiguity in family business mission statements. *Qualitative Research Reports in Communication*, 14(1), 87–96. doi: 10.1080/17459435.2013.835346
- Chomsky, N. (2002). An interview on minimalism. In A. Belletti & L. Rizzi (Eds.), *On nature and language* (p. 92-161). Cambridge: Cambridge University Press.
- Clark, A. (2016). *Surfing uncertainty: Prediction, action and the embodied mind*. Oxford, UK: Oxford University Press.
- Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In D. Noelle et al. (Eds.), *Proceedings of 37th Annual Meeting of the Cognitive Science Society*. Austin, TX.
- Evans, O., Stuhlmüller, A., & Goodman, N. (2016). Learning the preferences of ignorant, inconsistent agents. In *Thirtieth aaai conference on artificial intelligence*.
- Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation: Advances in Research and Theory*, 49, 209-246.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336, 998-998.
- Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, 11(5), e0154854.
- Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1), 3–44.
- Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 25–50.

- Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, 6, 187-214. doi: 10.1080/17588928.2015.1020053
- Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11), 818-829.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (p. 26-40). New York: Academic Press.
- Hawkins, R. X., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? good questions provoke informative answers. In D. R. W. A. S. Y. J. M. T. J. C. D. . M. P. P. Noelle D. C. (Ed.), *Proceedings of 37th annual meeting of the cognitive science society* (pp. 878–883). Austin, TX.
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8), 589–604.
- Jones, E. E., & Davis, K. E. (1965). From acts to dispositions the attribution process in person perception. In *Advances in experimental social psychology* (Vol. 2, pp. 219–266). Elsevier.
- Kao, J., Bergen, L., & Goodman, N. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society meeting of the cognitive science society*.
- Kelley, H. H. (1967). Attribution theory in social psychology. In *Nebraska symposium on motivation*.
- Kelley, H. H., & Stahelski, A. J. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of personality and social psychology*, 16(1), 66 – 91.
- Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.
- Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks: Sage Publications.
- Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, 9(3), e92160.
- Mohr, L. B. (1983). The implications of effectiveness theory for managerial practice in the public sector. In K. S. Cameron & D. A. Whetten (Eds.), *Organizational effectiveness* (pp. 225–239). Elsevier.
- Ossa-Richardson, A. (2019). *A history of ambiguity*. Princeton University Press.
- Pascale, R. T., & Athos, A. G. (1981). *The art of Japanese management*. New York: Simon & Schuster.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122, 280-291.
- Qing, C., & Franke, M. (2015). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In H. Zeevat & H.-C. Schmitz (Eds.), *Bayesian natural language semantics and pragmatics* (p. 201-220). Springer.



- Sennet, A. (2016). Ambiguity. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2016 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2016/entries/ambiguity/>.
- 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.
- Sikos, L., Venhuizen, N., Drenhaus, H., & Crocker, M. (2019, 04). *Reevaluating pragmatic reasoning in web-based language games*. doi: 10.13140/RG.2.2.30535.14249
- Wasow, T. (2015). Ambiguity avoidance is overrated. In S. Winkler (Ed.), *Ambiguity: Language and communication* (p. 29-47). de Gruyter.
- Woensdregt, M., Kirby, S., Cummins, C., & Smith, K. (2016). Modelling the co-development of word learning and perspective-taking. In *Cogsci*.
- Yoon, E. J., Frank, M. C., Tessler, M. H., & Goodman, N. D. (2018). Polite speech emerges from competing social goals.

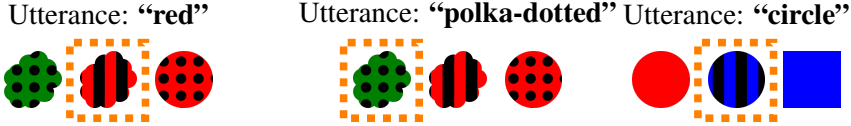
## A Ambiguity classes

### Experiment 1

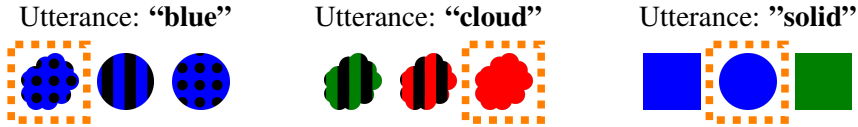
Figure 9 shows three exemplar scenarios for three representative ambiguity classes. Let us consider the first class in more detail. In the scenario  $S$  on the left side of Figure 9a, the utterance “blue” refers either to the blue square or the blue circle. The picked object, that is, the blue circle, is unique in its shape (circle) and shares the other non-referenced property with both other objects (that is, its plain pattern). The referenced but not picked object (that is, the blue square), shares its shape with the non-referenced object. In the scenario  $S$  in the center, the referenced two red objects differ in texture but share shape with the non-referenced object. In the scenario  $S$  on the right, the referenced two solid objects can be contrasted in their color but share their shape with the third object.



(a) The utterance references two objects, the picked object has one non-referenced unique feature, while the other, non-referenced feature is shared amongst all three objects. The other referenced, but not chosen object, shares its other feature with the non-referenced object.



(b) The utterance  $u$  references two objects whereby both objects only share the uttered feature. The third object shares one feature with each of the two referenced objects.



(c) In this third exemplar ambiguity class, the utterance refers to all three objects. The picked object shares one feature with one other object and has one feature just for itself while the other two objects share it.

Figure 9: Three exemplar scenarios  $S$ , constraining utterance  $u$ , and chosen object  $s$  are shown for three exemplar ambiguity classes for Experiment 1.

### Experiment 2

Figure 10 shows three exemplar scenarios for three representative ambiguity classes. Let us again consider the first class in more detail. In the scenario  $S$  on the left side of Figure 10a, all three objects share the feature pattern (solid), while two share the

color (blue), and the other two share the shape (square). As a result, uttering *green* or *circle* will give no choice to the listener because the utterance identifies one unique object. On the other hand, uttering *solid* will let the listener choose freely, while uttering *blue* or *square* will give a specific choice between two objects, that is, between the blue circle and the blue square or between the blue square or the green square, respectively. In the scenario *S* in the center, the objects share the shape (circle), two share the pattern (solid), and the other two share the color (red). Here, *circle* references all three objects, *red* or *solid* reference pairs of objects, and *striped* or *green* reference one unique object each. In the scenario *S* on the right, the object again share the shape (cloud), two share the pattern (solid), while the other two share the color (blue).



(a) In this exemplar ambiguity class, one feature is shared by all three objects, while the two other features allow the distinction between two different pairs of objects and the reference of one of two uniquely identifiable objects.



(b) In this second exemplar ambiguity class, all three feature types allow the identification of pairs of objects or unique objects, where all three features contain one unique feature type, each. As a result, there are three utterances that each pick out a different pair of objects and three other utterances that each reference one single object – effectively allowing the unique identification of each object as well as the identification of all three possible pairs.



(c) In this third exemplar ambiguity class, two features have three unique values, while one feature allows the identification of a pair of objects.

Figure 10: Three exemplar scenarios *S* are shown for three exemplar ambiguity classes for Experiment 2.