Semantics & Pragmatics Volume 4, Article 1: 1–82, 2011
doi: 10.3765/sp.4.1

Quantity implicatures, exhaustive interpretation, and rational conversation*

Michael Franke
University of Tübingen

Received 2010-10-30 / First Decision 2011-01-13 / Revision Received 2011-02-07 / Second Decision 2011-03-19 / Revision Received 2011-03-27 / Published 2011-06-21

Abstract Quantity implicatures are inferences triggered by an utterance based on what other utterances a speaker could have made instead. Using ideas and formalisms from game theory, I demonstrate that these inferences can be explained in a strictly Gricean sense as *rational behavior*. To this end, I offer a procedure for constructing the context of utterance insofar as it is relevant for quantity reasoning as a game between speaker and hearer. I then give a new solution concept that improves on classical equilibrium approaches in that it uniquely selects the desired "empirically correct" play in these interpretation games by a chain of back-and-forth reasoning about players' behavior. To make this formal approach more accessible to a wider audience, I give a simple algorithm with the help of which the model's solution can be computed without having to do heavy calculations of probabilities, expected utilities and the like. This rationalistic approach subsumes and improves on recent exhaustivity-based approaches. It makes correct and uniform predictions for quantity implicatures of various epistemic varieties, free choice readings of disjunctions, as well as a phenomenon tightly related to the latter, namely so-called "simplification of disjunctive antecedents".

Keywords: quantity implicature, exhaustive interpretation, game theory, iterated best response

* I am very grateful for countless conversations with many colleagues and friends that helped shape the thoughts of this paper. In particular, I would like to express my sincere gratitude to my teachers Robert van Rooij, Martin Stokhof and Gerhard Jäger for many invaluable lessons and other acts of kindness. Thanks also to Tikitu de Jager for highly-esteemed companionship, to Anton Benz for his support, and to Christian Ebert and Jason Quinley for help on the manuscript. The paper has benefited enormously from the critical but constructive comments provided by David Beaver and three anonymous referees. Remaining errors are mine.

1 Introduction

In his essay *Logic & Conversation* (Grice 1975), Paul Grice made a beautiful case for parsimony of a theory of meaning in a defense of logical semantics. Grice maintained that semantic extravagance is often unnecessary, and he showed that many alleged differences between classical logical semantics and intuition can be explained systematically as arising from certain regularities of our conversational practices, which he summarized under the label *Maxims of Conversation*. Based on the assumption that the speaker adheres to these maxims, the hearer is able to pragmatically enrich the semantic meaning of an utterance in systematic ways. This is, in fairly simplified terms, the background of Grice's theory of *conversational implicatures*.[1]

A particular kind of conversational implicature are so-called *quantity implicatures*. For illustration, consider the following conversation:

(1) BUBU: All of my friends are metalheads.
KIKI: Some of my friends are too.

From Kiki's reply in (1), we readily and reasonably infer the following:

(2) It's not the case that all of Kiki's friends are metalheads.

Does that mean that the quantifier *some* means "some but not all"? Absolutely not, Grice would say. The literal meaning of *some* may very well be just the existential quantifier familiar from first order logic. This is because we can explain the inference in (2) as something that Bubu, the hearer, is entitled to infer based on the assumption that Kiki, the speaker, is forthcoming and cooperative towards the interest shared by speaker and hearer, which is — so the central Gricean assumption goes — honest and reliable information exchange. In other words, Bubu may reasonably assume that Kiki's linguistic

1 Grice's theory of conversational implicatures was developed further in many ways by a great number of linguists and philosophers. I do not wish to claim that I am producing Grice's original view on the matter, but rather that which I take to be the distilled common idea that emerged and is still developed to-day in the community. Crucial steps in the shaping of our ideas about conversational implicatures were made by, among others, Horn (1972), Gazdar (1979), Atlas & Levinson (1981), Levinson (1983), Horn (1984), Horn (1989), and Levinson (2000).

behavior is governed by the following *Speaker Quantity Principle*:[2]

(3) Speaker Quantity Principle:
A cooperative speaker provides all true and relevant information she is capable of.

Based on this, Bubu may reason that it would have been more informative — as far as logical strength goes — for Kiki to simply have responded "All of mine too." The extra information would have been relevant, at least for the sake of small-talk conversation. Hence, the reason why Kiki has not simply said "All of mine too" was most likely so as not to speak untruthfully. So it must be the case that (2) is true.

This sketchy derivation clearly needs to be improved if it is to be the backbone of a respectable theory of quantity implicature *tout-court*. Firstly, it is clearly desirable to trade in the above informal reasoning for a more perspicuous, calculable and formal theory of quantity inferences, especially when we wish to assess predictions in more involved cases. This has been done, with ample degree of success (e.g., Gazdar 1979, Schulz & van Rooij 2006, Spector 2006). But it is also fair to say that the more formally rigorous a theory of quantity implicature is, the further removed it usually stands from the original idea behind the Gricean approach to conversational implicatures (cf., Geurts & Pouscoulous 2009b: 1–3). Which is what exactly?

Grice was an extremely careful philosopher, and as such he was naturally very aware of the tentativeness of his own proposal (cf., Chapman 2005). Indeed, Grice envisaged a *rationalistic foundation* of the Maxims of Conversation and the explanations of conversational implicatures that they license:

> "As one of my avowed aims is to see talking as a special case or variety of purposive, indeed rational, behaviour, it may be worth noting that the specific expectations or presumptions connected with at least some of the foregoing maxims have their analogues in the sphere of transactions that are not talk exchanges." (Grice 1975: 47)

2 This principle is not found in Grice's writings, but is a simplified and condensed conglomerate of Grice's Maxims of Quantity, Quality and Relevance. The present formulation does echo, however, rather closely the formulation given by Schulz & van Rooij (2006) in motivation of their exhaustivity operator (to be introduced in Section 4.1).

This essay is in this general spirit of Grice's programme. The main question to be asked is whether complex cases of quantity implicatures, can be grounded in rationalistic terms, using the formal machinery of game theory. After recapitulating some of the relevant issues concerning the notion of quantity implicature in Section 2, we will introduce a small test set of quantity implicatures in Section 3. Some of these cases are problematic even for our currently best formal theories, which I take to be theories based on the idea of *exhaustive interpretation*, to be recapitulated in Section 4. Section 5 then addresses the general question what a rationalistic explanation of quantity implicatures would have to look like. This paves the way for the game theoretic model to be introduced in Sections 6, 7 and 8. Section 9 discusses the model's predictions and Section 10 relates the present proposal to alternative approaches before Section 11 concludes. Two appendices provide more formal detail, one on exhaustive interpretation, one on game theoretic concepts.

The model that this paper introduces, called *iterated best response (*IBR*) model*, consists of two components (see Figure 1). The first component, subject of Section 6, is a principled procedure with which to construct a *game model* as a representation of the context of utterance, insofar as it is relevant to quantity reasoning. This part is often neglected in game theoretic approaches to pragmatics but absolutely crucial for a self-contained, falsifiable theory. This paper therefore shows how to construct a unique game from a to-be-interpreted target utterance and its alternative expressions, and attempts to trace each step in this construction to its underlying intuition about the nature of cooperative conversation.

The second component, given in Section 8, is a *solution concept* that captures the "pragmatic reasoning" in a given context game model. The solution concept that this paper spells out, called *iterated best response* (IBR) reasoning, builds on conceptually similar approaches taken, among others, by Benz (2006), Stalnaker (2006), Benz & van Rooij (2007) and Jäger & Ebert (2009). The solution concept of this paper demonstrably makes better predictions than previous approaches using standard equilibrium notions or refinements thereof (cf., Parikh 1991, 2001, de Jager & van Rooij 2007, van Rooij 2008). Unlike the latter, IBR reasoning makes transparent reference to interlocutors' mental states, making it more accessible to commonsense intuition (as opposed to game theoretic expertise). Pragmatic reasoning is spelled out in terms of back-and-forth reasoning about the players' behavior. Beginning with the assumption that utterances are true, interlocutors reason

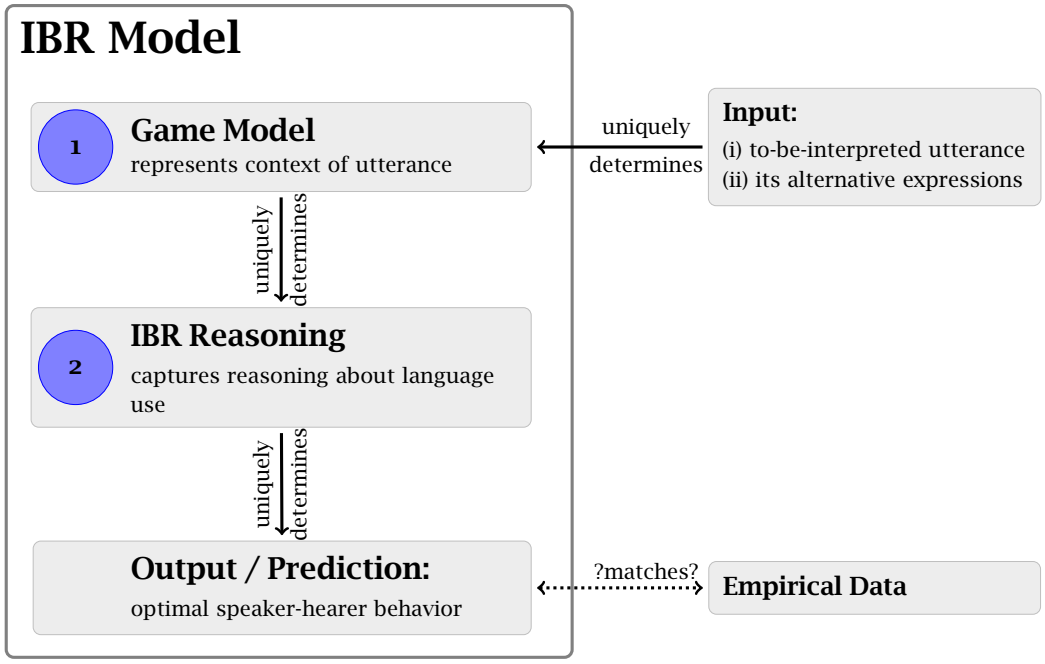


Figure 1 Architecture of the IBR model

pragmatically about which utterances and interpretations are rational given the belief that their conversational partner reason similarly. Repeated application of this reasoning scheme singles out uniquely strategic behavior that would ideally correspond to the empirically observed pragmatic language use.

More specifically, the IBR model's treatment of quantity implicatures is essentially as follows. Given an utterance m^* whose interpretation we are interested in and its potential alternatives $m_1 \dots m_n$, the game model considers a number of state distinctions, where a state is a set of all those possible worlds that agree on the truth values of expressions $m_1 \dots m_n$. The model assumes that the speaker but not the hearer knows the actual state, and that speaker and hearer are both interested in the hearer being able to deduce the actual state as precisely as possible. By IBR reasoning, then, each player, i.e, speaker and hearer, associates states and expressions by considering how states and messages are associated by the other player. Each association step aims to optimize communication, making it less ambiguous or error-prone. By iterating this procedure, the model selects a pair of

speaker and hearer behavior that is in equilibrium, i.e., that cannot be further improved by one player alone. The paper shows that equilibrium behavior that is selected for in this way explains a variety of quantity implicatures as mutually optimized language use.

In particular, the paper focuses on the implicatures of disjunctions, either non-embedded, or embedded under existential modals, or occurring within the antecedents of conditionals. It is an open question that this paper does not address in full detail whether the IBR model, as presented here, is capable of explaining all of the rather intricate data surrounding disjunctions in various other embedded positions (Chierchia, Fox & Spector 2008, 2009, Fox & Spector 2009).

The model proposed here is similar (but not in general equivalent) to a variant of iterated exhaustive interpretation, and can be conceived as providing a conceptual grounding thereof. This leads to a very surprising conclusion: non-iterated exhaustive interpretation and, more generally, simple kinds of quantity implicatures can be rationalized without appeal to a Gricean speaker maxim like (3), solely in terms of rational interpretation based on the given semantic meaning alone. This is spelled out in Section 10.

2 Quantity implicatures: Background

Epistemic variety. There is more variety in the notion of a quantity implicature than the discussion in the last section suggested. If we compare an utterance of (4) to an utterance of (5), this may actually give rise to any of the following varieties of a quantity implicature: the *general epistemic implicature* in (6a), the *strong epistemic implicature* in (6b), the *weak epistemic implicature* in (6c), or the *base-level quantity implicature* in (6d).

(4) Some of Kiki's friends are metalheads. "some"

(5) All of Kiki's friends are metalheads. "all"

(6) a. The speaker does not believe that all of Kiki's friends are metalheads. $\neg \text{Bel}_S$"all"

b. The speaker believes that not all of Kiki's friends are metalheads. $\text{Bel}_S \neg$"all"

c. The speaker is uncertain whether all of Kiki's friends are metalheads. Uc_S"all"

d. It's not the case that *all* of Kiki's friends are metalheads. $\neg$"all"

The general epistemic implicature in (6a) is usually considered most basic, because it is this inference only that we are strictly speaking entitled to draw when a principle like (3) tells us that the speaker was not able to truthfully assert the stronger statement (5). The strong epistemic implicature (6b) and the base-level implicature (6d) can then in fact be derived from the general epistemic implicature under an additional assumption of *sender competence*:[3] if we assume that the speaker knows how many of her friends are fans of heavy metal, then (6a) may be strengthened to (6b), which may in turn be strengthened to (6d). However, it is often overlooked that the general epistemic implicature in (6a) can also be strengthened the other way (see Geurts 2010: §2.3): if we assume that the speaker is *in*competent, i.e., that she is likely *not* informed properly, then we can strengthen (6a) to yield the weak epistemic implicature in (6c) too.[4]

Expression alternatives. Which quantity implicatures can be derived obviously hinges on which expression alternatives we consider. If we had considered also the logically stronger expression:

(7) Some but not all of Kiki's friends are metalheads.

derivation of the base-level implicature in (6d) would have been jeopardized; in that case, we would only derive that the speaker must be uncertain whether (5) or (7). This is why, traditionally, *expression alternatives* were taken to be derived by lexical substitution of elements that are associated on an ordered *scale* of alternatives (e.g., Horn 1972, 1984). Quantity implicatures derived with the help of such a scale are therefore often referred to as *scalar implicatures*. To exclude pathological predictions, these scales were required to satisfy certain conditions, such as that elements had to be equally lexicalized, should not be more complex than the original utterance, or had

3 There are several conceptually and formally different implementations of this idea in the literature (cf., van Rooij & Schulz 2004, Sauerland 2004, Schulz 2005, Schulz & van Rooij 2006, Russell 2006, Geurts 2010).

4 The general logic of belief I adopt here is the standard possible-worlds analysis of Hintikka. If φ is a proposition (a set of possible worlds), then $\mathrm{Bel}_S\varphi$ expresses that the speaker believes that φ is true. This is the case, as is classically assumed, just when all the possible worlds which the speaker considers possible are contained in φ. I use notation $\mathrm{Uc}_S\varphi$ to denote that the speaker is uncertain whether φ is true, which is an abbreviation of $\neg\mathrm{Bel}_S\varphi \wedge \neg\mathrm{Bel}_S\neg\varphi$. With this there are then three *belief values*, as I will call them, i.e., three possibilities for an agent's beliefs to relate to a proposition: (i) the agent believes the proposition is true, (ii) believes that it is false, or (iii) is uncertain about it.

to have the same monotonicity properties (e.g., Atlas & Levinson 1981, Horn 1989, Matsumoto 1995). Some authors have since generalized the notion of an ordered scale to that of an unordered *set* of lexical alternatives (e.g., Sauerland 2004, Fox 2007), others have been even more liberal, or so it seems, by comparing utterances to the set of possible answers to some possibly implicit question under discussion (van Kuppevelt 1996, van Rooij & Schulz 2004, Schulz & van Rooij 2006, Spector 2006). The issue of which alternatives to consider for derivation of quantity implicatures is, however, still ongoing (cf., Katzir 2007, Swanson 2010).

I have no new theory to offer on this question. The system I am going to introduce in the following sections actually is not a theory of alternatives, but one of reasoning *about* alternatives: given a set of alternatives as input, it will specify which implicatures we may expect. Still, the game theoretic perspective that I am going to defend in the following does indirectly add to the debate. By explicitly modeling a speaker and a hearer and their beliefs about each other's beliefs and action alternatives, we are reminded of the fact that a conversational implicature is not something that attaches to a sentence in isolation. A conversational implicature, in the original Gricean sense, can only be derived from an utterance (cf., Bach 2006). If this is right, then the hearer's choice of which alternatives to consider must be made in context, based on assumptions about which alternatives the speaker may have been aware of in the first place, and which of those she considered worthwhile entertaining.[5] But that means that factors such as strength of lexical association, morphosyntactic complexity, and lexical semantic properties may all play *a* role in the hearer's construction of his assessment of the context of utterance, but it is unlikely, to say the least, that any *one* factor should be solely decisive (see Swanson 2010 for a similar opinion). In the following I will therefore largely rely on an intuitive understanding of what may count as a consistent, yet *prima facie* plausible set of alternative expressions in a given context.

3 Quantity implicatures: Some relevant cases

This paper will be mainly concerned with the interpretation of three types of disjunctive constructions: (i) *plain disjunctions* of the form $A \vee B$, (ii) *free*

5 This can be taken to higher-order reasoning as well, for the hearer could also exclude an alternative from consideration because he believes that, although the speaker is aware of it, the speaker believes that the hearer is not, etc.

choice disjunctions of the form $\Diamond(A \vee B)$ and (iii) disjunctions in antecedents of conditionals such as $(A \vee B) > C$. This section summarizes what is worth explaining about these constructions. The upshot of this discussion is also summarized in Figure 2 at the end of this section. As an additional test case, we should also consider the case of "some" and "all", as outlined above.

Disjunction. An utterance of a plain disjunction as in (8) is usually compared to the alternatives in (9) (cf., Sauerland 2004, Fox 2007), and it is associated with the *ignorance implicature* in (10) that the speaker is uncertain about each disjunct.

(8)	Martha is in love with Alf or Bert.	$A \vee B$
(9)	a. Martha is in love with Alf.	A
	b. Martha is in love with Bert.	B
	c. Martha is in love with Alf and Bert.	$A \wedge B$
(10)	a. The speaker is uncertain whether Martha is in love with Alf.	$\mathrm{Uc}_S A$
	b. The speaker is uncertain whether Martha is in love with Bert.	$\mathrm{Uc}_S B$

Additionally, an utterance of (8) may give rise to an *exclusivity implicature* as in (11), which may arise as a base-level or either of the three varieties of epistemic implicatures.

(11)	a. The speaker does not believe that Martha is in love with both Alf and Bert.	$\neg\mathrm{Bel}_S(A \wedge B)$
	b. The speaker believes that Martha is not in love with both Alf and Bert.	$\mathrm{Bel}_S\neg(A \wedge B)$
	c. The speaker is uncertain whether Martha is in love with both Alf and Bert.	$\mathrm{Uc}_S(A \wedge B)$
	d. Martha is not in love with both Alf and Bert.	$\neg(A \wedge B)$

However, it is not uncontroversial whether the exclusivity implicature should be traced unequivocally to quantity reasoning. It is conceivable that exclusivity implicatures arise due to world knowledge, such as when disjuncts are logically inconsistent, or when conjoined truth is highly unlikely (cf., Geurts 2010). I would like to remain neutral on this point, and I will consider both including a conjunctive alternative (9c) and not including it. As is to be expected, when including (9c), the system that later sections will introduce will

be able to derive both ignorance and exclusivity implicatures as a quantity inference; when we do not include (9c), we can only derive the ignorance implicatures.

Free choice disjunction. Intuitively, an utterance of a sentence like (12a) where a disjunction scopes under an existential modal operator has a so-called *free choice implicature* as in (12b).[6]

(12) a. You may take an apple or a pear. $\Diamond(A \vee B)$
b. You may take an apple and you may take a pear. $\Diamond A \wedge \Diamond B$

The inference from (12a) to (12b) is not valid under a standard logical semantics which treats disjunction as the usual Boolean connective and the modal as an existential quantifier over accessible worlds. In a Gricean spirit, we would like to account for this inference as a conversational implicature, indeed a base-level quantity implicature.[7] One of the arguments in favor of an implicature-based analysis is the observation that the inference in (12b) seems to rest on the contextual assumption that the speaker is, in a sense, an authority about the deontic modality in question. If this assumption is not warranted or explicitly suspended as in example (13) we get the *ignorance implicatures* in (14).

(13) You may take an apple or a pear, but I don't know which.

(14) a. The speaker is uncertain whether the hearer may take an apple. $\mathrm{Uc}_S(\Diamond A)$
b. The speaker is uncertain whether the hearer may take a pear. $\mathrm{Uc}_S(\Diamond B)$

Moreover, an utterance of (12a) may also receive *exclusivity implicatures* of

6 The *paradox of free choice permission* — so-called by von Wright (1968) — is a well-known problem in deontic logic (cf., Ross 1944, von Wright 1951). Influential solutions were attempted by Kamp (1973, 1978), but, more recently, the puzzle has inspired many more linguists to try a solution. Semantically oriented approaches reconsider the semantics of disjunctions (Zimmermann 2000, Geurts 2005, Simons 2005) or of the modals involved (Merin 1992, van Rooij 2000, Asher & Bonevac 2005, Barker 2010). Pragmatic approaches, more or less close in spirit to Grice's programme, have been given by others (Kratzer & Shimoyama 2002, Chierchia 2004, Schulz 2005, Fox 2007).

7 See also Kratzer & Shimoyama (2002), Alonso-Ovalle (2005), Schulz (2005) for more arguments why the free choice inference in (12a) should be treated as an implicature.

various strengths, as in (15).

(15) a. The speaker does not believe that the hearer may take both. $\neg \mathrm{Bel}_S \Diamond (A \wedge B)$
b. The speaker believes that the hearer may not take both. $\mathrm{Bel}_S \neg \Diamond (A \wedge B)$
c. The speaker is uncertain whether the hearer may take both. $\mathrm{Uc}_S \Diamond (A \wedge B)$
d. The hearer may not take both. $\neg \Diamond (A \wedge B)$

Again, as with plain disjunctions, I would like to remain neutral as to whether exclusivity implicatures always arise and should systematically be derived as a quantity implicature by comparison with a conjunctive alternative. If I tell my students:

(16) You may consult your notes or your textbook during the exam.

it seems to me that it requires emphatic stress on "or" to convey that consulting both notes and textbook is not an option.

This ties in with the question of which alternatives are operative in the derivation of these implicatures. As before with plain disjunctions, I will assume that an utterance of (12a) invites comparison with at least (17a) and (17b), and possibly, but not necessarily also (17c).

(17) a. You may take an apple. $\Diamond A$
b. You may take a pear. $\Diamond B$
c. You may take an apple and a pear. $\Diamond (A \wedge B)$

Simplification of disjunctive antecedents. The last case that I would like to consider concerns the interpretation of disjunction in the antecedents of conditionals. I will argue that we should preferably treat these in parallel with free choice readings of disjunctions under existential modals (cf., Klinedinst 2006).

Intuitively, the indicative conditional in (18) seems to convey both (19a) and (19b), and analogously the counterfactual conditional (20) seems to

convey both (21a) and (21b).[8]

(18) If you eat an apple or a pear, you will feel better. $(A \vee B) \boxRightarrow C$
(19) a. If you eat an apple, you will feel better. $A \boxRightarrow C$
b. If you eat a pear, you will feel better. $B \boxRightarrow C$
(20) If you'd eaten an apple or a pear, you'd feel better. $(A \vee B) \boxRightarrow C$
(21) a. If you'd eaten an apple, you'd feel better. $A \boxRightarrow C$
b. If you'd eaten a pear, you'd feel better. $B \boxRightarrow C$

Similar inferences are warranted for conditionals with existential modals in their consequents: an utterance of (22) may convey both (23a) and (23b); an utterance of (24) may convey both (25a) and (25b).

(22) If you eat an apple or a pear, you might feel better. $(A \vee B) \diamondRightarrow C$
(23) a. If you eat an apple, you might feel better. $A \diamondRightarrow C$
b. If you eat a pear, you might feel better. $B \diamondRightarrow C$
(24) If you'd eaten an apple or a pear, you might feel better. $(A \vee B) \diamondRightarrow C$
(25) a. If you'd eaten an apple, you might feel better. $A \diamondRightarrow C$
b. If you'd eaten a pear, you might feel better. $B \diamondRightarrow C$

In general, the inference from $(A \vee B) > C$ to $A > C$ (or $B > C$) is known as *simplification of disjunctive antecedents*, henceforth SDA.

Although SDA is a valid inference under a material implication analysis of conditionals, standard possible-worlds semantics in the vein of Stalnaker (1968) and Lewis (1973) do not make SDA valid. According to these theories, we evaluate a conditional as true or false in a given possible world w with respect to a set R_w of worlds accessible from w and some ordering $\preceq_w$ on this set. Many reasonable constraints on the nature of this ordering could be given to instantiate certain influential theories of conditionals (cf., Stalnaker 1968, Lewis 1973, Kratzer 1981, Veltman 1985). For the present pragmatic purpose we should remain noncommittal and not take on *any* particular constraints on the ordering, except that it be well-founded, i.e., that it conforms to the *limit assumption*, so as to facilitate notation. We then first define

(26) $\text{Min}(R_w, \preceq_w, A) = \{v \in R_w \cap A \mid \neg\exists v' \in R_w \cap A\colon\ v' \prec_w v\}$

8 I make use of the following notation: $A \boxRightarrow B$ is an abbreviation for "if A, then will/would B" and $A \diamondRightarrow B$ stands for "if A, then may/might B", irrespective of whether the conditional is in the subjunctive of the indicative mood. I write $A > B$ for any conditional with either universal or existential modal in the consequent.

and then define for indicative or counterfactual conditionals alike that:

(27) $A \boxRightarrow C$ is true in w iff $\mathrm{Min}(R_w, \preceq_w, A) \subseteq C$,

as well as:

(28) $A \diamondRightarrow C$ is true in w iff $\mathrm{Min}(R_w, \preceq_w, A) \cap C \neq \emptyset$.

Clearly, SDA is not a generally valid inference under these semantics, because, for instance, $(A \vee B) \boxRightarrow C$ could be true if all minimal worlds where $A \vee B$ is true are such that A and C are true and B is false, while all minimal B-worlds are worlds where C is false. In that case $(A \vee B) \boxRightarrow C$ would be true but $B \boxRightarrow C$ would be false. Whence that SDA is not semantically valid.

Should we worry? Some say yes, some say no. From those who say yes, the invalidity of SDA has been held as a problem case against in particular Lewis's (1973) theory of counterfactuals (see Nute 1975, Fine 1975), but the case would equally apply to indicatives under like-minded semantic theories. On the other hand, there are good arguments not to want SDA to be a semantically valid inference pattern. Warmbrōd (1981) gives a strong argument in favor of this position. He argues that if a conditional semantics makes SDA valid, and if we otherwise stick to standard truth-functional interpretation of disjunction, we can also derive that inferences like that from (29a) to (29b) are generally valid, which intuitively should not be the case.[9]

(29) a. If you eat an apple, you will feel better. $A \boxRightarrow C$
b. If you eat an apple and a rock, you'll feel better. $(A \wedge B) \boxRightarrow C$

Another argument against a semantic validation of SDA comes from examples such as the following (cf., McKay & van Inwagen 1977):

(30) a. If John had taken an apple or a pear, he would have taken an apple.
b. If John had taken a pear, he would have taken an apple.

If SDA was semantically valid then (30a) would imply (30b), but this is of course nonsense. Together, this suggests loosely that SDA should perhaps be

9 Formally, this is because if SDA is generally valid, we can infer from $A \boxRightarrow C$ and the fact that $(A \wedge B) \vee (A \wedge \neg B)$ is a truth-functionally equivalent to A that $((A \wedge B) \vee (A \wedge \neg B)) \boxRightarrow C$. Then, by SDA, we derive $(A \wedge B) \boxRightarrow C$ for arbitrary B.

thought of as a pragmatic inference on top of a standard semantics, which is what I argue for here.[10]

Indeed, it has been noted before that SDA looks strikingly similar to free choice permission in several respects (cf., Klinedinst 2006, van Rooij 2006). Firstly, English conditionals like (31) *can* be used to grant permissions.

(31) It's fine with me if you take an apple.

This may not be the most frequent and natural way of giving permissions, but it seems that at least the question

(32) Is it okay if I take an apple?

is an expression frequently used to *ask* for permission in English. This is different in other languages though. For instance, lacking a clear equivalent to English modal "may", a standard construction for permission giving in Japanese is the conditional construction "-te mo" which generally translates as "even if" (see McClure 2000: 180):

(33)

ringo	wo	tabe-te	mo	ii.
apple	Object Marker	eat-TE-Form	also	good

'It's good even if you eat an apple.'
'You may eat an apple.'

Most crucially, a conditional similar to (31) with a disjunctive antecedent as in (34) is certainly taken to convey that both sentences in (35) are true.

(34) It's fine with me if you take an apple or a pear.

(35) a. It's fine with me if you take an apple.
b. It's fine with me if you take a pear.

Moreover, although possibly marginal, it does seem that there are also epistemic ignorance readings for conditionals with disjunctive antecedents, quite parallel to cases like (13)–(14):

(36) a. If you eat an apple or a pear, you will feel better, but I don't know which. $(A \vee B) \boxdot\!\!\Rightarrow C$
b. The speaker is uncertain whether $A \boxdot\!\!\Rightarrow C$ is true. $\mathrm{Uc}_S(A \boxdot\!\!\Rightarrow C)$
c. The speaker is uncertain whether $B \boxdot\!\!\Rightarrow C$ is true. $\mathrm{Uc}_S(B \boxdot\!\!\Rightarrow C)$

10 So here I depart slightly from ultra-orthodox Griceanism, if you want to call it that: I consider (something like) the above order-sensitive possible worlds semantics the standard, and I do not consider material implication a respectable semantic analysis of conditionals.

construction	implicature	
plain disjunction	ignorance implicature:	$\mathrm{Uc}_S A$ & $\mathrm{Uc}_S B$
$A \vee B$	exclusivity implicature …	
	… general epistemic:	$\neg \mathrm{Bel}_S(A \wedge B)$
	… strong epistemic:	$\mathrm{Bel}_S \neg (A \wedge B)$
	… weak epistemic:	$\mathrm{Uc}_S(A \wedge B)$
	… base-level:	$\neg(A \wedge B)$
FC-disjunction	FC-implicature:	$\Diamond A \wedge \Diamond B$
$\Diamond(A \vee B)$	ignorance implicature:	$\mathrm{Uc}_S \Diamond A$ & $\mathrm{Uc}_S \Diamond B$
	exclusivity implicature …	
	… general epistemic:	$\neg \mathrm{Bel}_S \Diamond (A \wedge B)$
	… strong epistemic:	$\mathrm{Bel}_S \neg \Diamond (A \wedge B)$
	… weak epistemic:	$\mathrm{Uc}_S \Diamond (A \wedge B)$
	… base-level:	$\neg \Diamond (A \wedge B)$
SDA-disjunction	SDA-implicature:	$(A > C) \wedge (B > C)$
$(A \vee B) > C$	ignorance implicature:	$\mathrm{Uc}_S(A > C)$
		& $\mathrm{Uc}_S(B > C)$
	exclusivity implicature …	
	… general epistemic:	$\neg \mathrm{Bel}_S((A \wedge B) > C)$
	… strong epistemic:	$\mathrm{Bel}_S \neg ((A \wedge B) > C)$
	… weak epistemic:	$\mathrm{Uc}_S((A \wedge B) > C)$
	… base-level:	$\neg((A \wedge B) > C)$

Figure 2 Overview of quantity implicatures that this paper focuses on

If the speaker's epistemic uncertainty is explicitly marked, we infer from (36a) that the speaker is uncertain whether (36b), respectively (36c) is true.

Taken together, SDA looks remarkably similar to free choice implicatures and we should try to see whether both can be explained uniformly as a quantity implicature (cf., Klinedinst 2006 for a rather different uniform explanation of these phenomena). For overview, the table in Figure 2 lists once more all of the constructions and inferences that we will be concerned with in the remainder of this paper. For comparison, the following section looks at some of the predictions of exhaustification-based approaches for these test cases.

4 Exhaustive interpretation

The currently most prominent formal approaches to quantity implicatures are formulated in terms of *exhaustivity operators*. These were originally conceived to account for the exhaustive reading of answers to questions (cf., Groenendijk & Stokhof 1984, von Stechow & Zimmermann 1984). When it comes to quantity implicatures, there are two conceptually and formally distinct versions of exhaustivity operators, one is more semantic, the other is more syntactic in nature. The semantic approach tries to account for quantity implicature as interpretation in *minimal models* (cf., van Rooij & Schulz 2004, Schulz & van Rooij 2006, Spector 2006). This is in contrast to an approach due to Fox (2007) in terms of a notion called *innocent exclusion* (see also Alonso-Ovalle 2008, Chierchia et al. 2008). The latter is not formulated in terms of possible worlds, but rather in terms of explicit reasoning about alternatives. That is why I refer to it as a syntactic approach. The following briefly introduces both of these approaches in sufficient formal detail for further comparison.

4.1 The minimal-models approach

The general idea of the minimal-models approach to exhaustive interpretation of a sentence S is to define an ordering on possible worlds (or the speaker's information states) based on the available alternatives ALT of S (with $S \in$ ALT) and to consider as pragmatic interpretation of a given sentence all the worlds (or states) that are minimal with respect to this ordering.[11] The relevant ordering is obtained from ALT by defining a world (or state) to be more minimal than another if strictly more alternatives from ALT are false (not believed true). Its proponents suggest that in this way the minimal-models approach to exhaustive interpretation captures pragmatic interpretation based on (something like) the Speaker Quantity Principle (3), because the interpretation selects those worlds or states where fewest possible alternatives are true, because, in turn, a cooperative speaker may be expected to have uttered these alternatives, if she had had the relevant knowledge. (Section 10 gives a surprising alternative characterization of this operation as rational hearer

11 Here and in the following, I will be sloppy in using capital letters as variables for both sentences and the propositions (sets of possible worlds) that they denote. To make matters *even worse*, I happily administer logical notation to the mix: for instance, $\neg S$ would be the negation of sentence S, or the complement of set S in the set of all possible worlds.

interpretation based on a presumption of truthfulness alone.)

Base-level implicatures. To account for base-level implicatures, we consult a partial ordering on possible worlds $<_{\text{ALT}}$, defined as follows:

(37) $w' <_{\text{ALT}} w \qquad \text{iff} \qquad \{A \in \text{ALT} \mid w' \in A\} \subset \{A \in \text{ALT} \mid w \in A\}$.

The base-level exhaustive interpretation of S is then obtained from the assumption that the actual world is minimal with respect to this order in the set of worlds where S is true. This is captured in the following operator for exhaustive interpretation in terms of minimal models:

(38) $\text{EXH}_{\text{MM}}(S, \text{ALT}) = \{w \in S \mid \neg\exists w' \in S\colon\ w' <_{\text{ALT}} w\}$.

For example, take a contrast between sentences, "some A's are B's" and "all A's are B's" (such as between (4) and (5)). Then, assuming that the target utterance "some A's are B's" is true, we need to consider only two types of possible worlds: $w_{\exists\neg\forall}$ where some but not all A's are B's, and $w_{\forall}$ where all A's are B's. The ordering defined above gives us: $w_{\exists\neg\forall} <_{\text{ALT}} w_{\forall}$, because in $w_{\forall}$ both alternative sentences are true, while in $w_{\exists\neg\forall}$ only the weaker "some" statement is. Consequently, the pragmatic interpretation according to this approach selects:

(39) $\text{EXH}_{\text{MM}}(\text{"some } A\text{'s are } B\text{'s"}, \text{ALT}) = \{w_{\exists\neg\forall}\}$.

Notice that $w_{\exists\neg\forall}$ is a stand-in for a class of possible worlds: the pragmatic interpretation of the sentence is that some but not all A's are B's, just as we want it to.

Epistemic implicatures. To account for epistemic implicatures in terms of minimal models, we consult an ordering $<^{\Box}_{\text{ALT}}$ defined not on possible worlds, but on information states of the speaker.[12] As usual, an information state s is a non-empty set of possible worlds, collecting the possibilities that cannot be ruled out for certain by, in our case, the speaker. To compute general epistemic implicatures, information states are compared with respect to how many propositions from ALT the speaker knows to be true:

(40) $s' <^{\Box}_{\text{ALT}} s \qquad \text{iff} \qquad \{A \in \text{ALT} \mid s' \subseteq A\} \subset \{A \in \text{ALT} \mid s \subseteq A\}$.

12 My exposition here follows mainly the variant by van Rooij & Schulz (2006).

Based on this ordering, we define exhaustive interpretation that yields general epistemic implicatures as follows:

(41) $\text{Exh}^{\text{GE}}_{\text{MM}}(S, \text{Alt}) = \{s \subseteq S \mid \neg\exists s' \subseteq S\colon\ s' <^{\Box}_{\text{Alt}} s\}$.

This operator collects all the information states where fewest alternatives are *believed true* within the class of information states where the target sentence S is believed true.

To see how this accounts for the general epistemic quantity implicature of cases like the comparison between "some A's are B's" and "all A's are B's", we need to consider all information states expressible in terms of the types of worlds $w_{\exists\neg\forall}$ and $w_{\forall}$ from above. These are: $s_1 = \{w_{\exists\neg\forall}, w_{\forall}\}$ where the speaker is uncertain, $s_2 = \{w_{\exists\neg\forall}\}$ where the speaker believes that some but not all A's are B's, and $s_3 = \{w_{\forall}\}$ where the speaker believes that all A's are B's. The ordering $<^{\Box}_{\text{Alt}}$ as defined above yields: $s_1, s_2 <^{\Box}_{\text{Alt}} s_3$. With this, we derive:

(42) $\text{Exh}^{\text{GE}}_{\text{MM}}(\text{"some } A\text{'s are } B\text{'s"}, \text{Alt}) = \{s_1, s_2\}$,

which says that the speaker is in an information state where she does not believe that all A's are B's.

This prediction can be strengthened to obtain strong epistemic implicatures if we assume that the speaker is competent. Following van Rooij & Schulz, this can be expressed by layering another ordering $<^{\Diamond}_{\text{Alt}}$ on the outcome of interpretation $\text{Exh}^{\text{GE}}_{\text{MM}}(S, \text{Alt})$. This ordering $<^{\Diamond}_{\text{Alt}}$ ranks information states with respect to which alternatives from A are considered possible:

(43) $s' <^{\Diamond}_{\text{Alt}} s$ iff $\{A \in \text{Alt} \mid s' \cap A \neq \emptyset\} \subset \{A \in \text{Alt} \mid s \cap A \neq \emptyset\}$.

Using this ordering, we can define an exhaustivity operator to capture strong epistemic implicatures as follows:

(44) $\text{Exh}^{\text{SE}}_{\text{MM}}(S, \text{Alt}) = \{s \in \text{Exh}^{\text{GE}}_{\text{MM}}(S, \text{Alt}) \mid \neg\exists s' \in \text{Exh}^{\text{GE}}_{\text{MM}}(S, \text{Alt})\colon\ s' <^{\Diamond}_{\text{Alt}} s\}$.

In the example above, we get $s_2 <^{\Diamond}_{\text{Alt}} s_1$, so that:

(45) $\text{Exh}^{\text{SE}}_{\text{MM}}(\text{"some } A\text{'s are } B\text{'s"}, \text{Alt}) = \{s_2\}$.

The strong epistemic implicature that the speaker believes that some but not all A's are B's is predicted. (The minimal-models approach does not attend in detail to the difference between general and weak epistemic implicatures.)

4.2 The innocent-exclusion approach

Fox (2007) offers a different approach to exhaustive interpretation to account for base-level quantity implicatures.[13] Fox's approach makes use of a novel notion which he calls *innocent exclusion*. The idea is that quantity implicatures of S are computed by negating all alternatives that can be excluded *consistently* without making an *arbitrary* choice in excluding these. Towards a definition of innocent exclusion, first define what it means to be *consistently excludable*. A subset A of ALT is consistently excludable if negating all elements in A is consistent with the truth of S:

$$\text{(46)} \quad \text{CE}(S, \text{ALT}) = \left\{ X \subseteq \text{ALT} \mid \bigwedge_{A \in X} \neg A \text{ is consistent with } S \right\} .$$

We would like to exclude as many of the alternatives as possible. So we should look at the set of maximal elements in $\text{CE}(S, \text{ALT})$:

$$\text{(47)} \quad \text{Max-CE}(S, \text{ALT}) = \{X \in \text{CE}(S, \text{ALT}) \mid \neg \exists Y \in \text{CE}(S, \text{ALT}) \colon X \subset Y\} .$$

So, for each set of alternatives $X \in \text{Max-CE}(S, \text{ALT})$, negating all elements $A \in X$ would be a maximally consistent pragmatic enrichment of the target sentence S. But it may be the case that some alternatives $A \in \text{ALT}$ occur only in some but not all of the maximal sets in $\text{Max-CE}(S, \text{ALT})$. Excluding these would be an *arbitrary* choice. That's why Fox defines the set of innocently excludable alternatives $\text{IE}(S, \text{ALT})$ to S given ALT as those alternatives that are in *every* maximally consistent pragmatic enrichment:

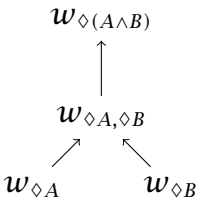
$$\text{(48)} \quad \text{IE}(S, \text{ALT}) = \bigcap \text{Max-CE}(S, \text{ALT}) .$$

Exhaustive interpretation based on innocent exclusion is then defined as:

$$\text{(49)} \quad \text{EXH}_{\text{IE}}(S, \text{ALT}) = S \wedge \bigwedge_{A \in \text{IE}(S, \text{ALT})} \neg A .$$

How does this definition differ from the minimal-models approach? Generally the operators EXH_{MM} and EXH_{IE} are not equivalent. The precise formal relation is worked out in Appendix A. In a nutshell, exhaustification based on innocent exclusion is always subsumed under the minimal models interpretation, but that the latter may rule out worlds that the former doesn't,

13 According to Fox, base-level quantity implicatures are computed as part of the syntactic system by his exhaustive operator, while epistemic implicatures are computed in the more traditional Gricean manner in the vein of Sauerland (2004).



	$\Diamond A$	$\Diamond B$	$\Diamond(A \wedge B)$	$\Diamond(A \vee B)$
$w_{\Diamond(A\wedge B)}$	1	1	1	1
$w_{\Diamond A,\Diamond B}$	1	1	0	1
$w_{\Diamond A}$	1	0	0	1
$w_{\Diamond B}$	0	1	0	1

Figure 3 Ordering on worlds for base-level free choice implicature

depending on the given set of alternatives. This property also matters for explanations of the basic free choice implicature which, for illustration, I will briefly go through next.

Predictions for basic free choice disjunctions. Recall that we are interested in explaining how (12a) can have the base-level implicature (12b), repeated here.

(12a) You may take an apple or a pear. $\Diamond(A \vee B)$
(12b) You may take an apple and you may take a pear. $\Diamond A \wedge \Diamond B$

If we take the set of alternatives $\text{ALT} = \{\Diamond A,\ \Diamond B,\ \Diamond(A \wedge B),\ \Diamond(A \vee B)\}$, then, in order to calculate the predictions of the minimal models approach, we have to distinguish four kinds of worlds and consider the partial ordering $<_{\text{ALT}}$ on them, as depicted in Figure 3. The table gives the truth values of the alternatives in each type of world. An arrow from w to w' symbolizes $w <_{\text{ALT}} w'$. (Arrows that follow from transitivity are left out for readability.) The minimal worlds according to that ordering are:

(50) $\text{EXH}_{\text{MM}}(\Diamond(A \vee B), \text{ALT}) = \{w_{\Diamond A}, w_{\Diamond B}\}$.

This prediction is too strong, for it actually rules out the free choice implicature in (12b). The free choice implicature would correspond to an interpretation that selects $\{w_{\Diamond A,\Diamond B}\}$ as the pragmatic interpretation for our target sentence.

Innocent exclusion is weaker on this set of alternatives and, crucially, does *not* exclude the free choice reading. In this case, there are two maximal sets of consistently excludable alternatives:

(51) $\text{Max-CE}(\Diamond(A \vee B), \text{ALT}) = \{\{\Diamond A,\ \Diamond(A \wedge B)\},\ \{\Diamond B,\ \Diamond(A \wedge B)\}\}$.

The intersection of these contains only $\Diamond(A \wedge B)$, so that:

(52) $\textsc{Exh}_{\textsc{IE}}(\Diamond(A \vee B), \textsc{Alt}) = \{w_{\Diamond A}, w_{\Diamond B}, w_{\Diamond A, \Diamond B}\}$.

In sum, $\textsc{Exh}_{\textsc{IE}}$ does not rule out the free choice inference, but it does not predict it either.

Iterated exhaustification. This is why Fox (2007) suggests applying the exhaustification operator, if needed, several times, so as also to factor in the exhaustive interpretation of the alternatives. The repeated application of exhaustification based on innocent exclusion is defined as follows:[14]

(53)
$$\begin{aligned} \textsc{Alt}_1 &= \textsc{Alt} \\ \textsc{Exh}_1(A) &= \textsc{Exh}_{\textsc{IE}}(A, \textsc{Alt}_1) \\ \textsc{Alt}_{n+1} &= \{\textsc{Exh}_n(A) \mid A \in \textsc{Alt}_n\} \\ \textsc{Exh}_{n+1}(A) &= \textsc{Exh}_{\textsc{IE}}(\textsc{Exh}_n(A), \textsc{Alt}_{n+1}) \,. \end{aligned}$$

Fox shows that further iteration of the exhaustification operator derives the basic free choice reading. I briefly repeat Fox's result here. Establish first the non-iterated exhaustive readings of all alternatives:

(54)
$$\begin{aligned} \textsc{Exh}_1(\Diamond A) &= \Diamond A \wedge \neg\Diamond B \\ \textsc{Exh}_1(\Diamond B) &= \Diamond B \wedge \neg\Diamond A \\ \textsc{Exh}_1(\Diamond(A \wedge B)) &= \Diamond(A \wedge B) \\ \textsc{Exh}_1(\Diamond(A \vee B)) &= \Diamond(A \vee B) \wedge \neg\Diamond(A \wedge B) \,. \end{aligned}$$

This is then the new set of alternatives $\textsc{Alt}_2$ for input into another round of exhaustification. This will affect only the target sentence, whose new interpretation becomes:

(55)
$$\begin{aligned} \textsc{Exh}_2(\Diamond(A \vee B)) &= \textsc{Exh}_1(\Diamond(A \vee B)) \wedge \neg\textsc{Exh}_1(\Diamond A) \wedge \neg\textsc{Exh}_1(\Diamond B) \\ &= \{w_{\Diamond A, \Diamond B}\} \,. \end{aligned}$$

Voila, free choice derived!

14 Spector (2007) utilises a parallel definition of iterated exhaustification in terms of the minimal models approach. It is easy to verify that this will not help to derive the basic free choice implicature, as the desired reading is already excluded at the first step and moreover, $\textsc{Exh}_{n+1}(A) \subseteq \textsc{Exh}_n(A)$ for all A and n.

It is worthwhile to mention that the procedure has already reached a *fixed point*: all subsequent iterations will yield the same outcome. It is easy to check, and in fact Spector (2007) provides a proof for his version of iterated exhaustification based on minimal models, that for finite set ALT the system must reach a fixed point after finitely many iteration steps. Essentially, this follows from the simple observation that iterated exhaustification is a monotonic operation: $\text{EXH}_{n+1}(A) \subseteq \text{EXH}_n(A)$ for all n.

A problem for iterated exhaustification. Fox's proposal neatly accounts for the basic free choice reading of sentences like (12a), but it does have its problems too. Despite the superficial parallel between free choice implicatures and SDA, Fox's system does not predict SDA, if we assume a standard Lewis-Stalnaker semantics for conditionals as outlined in Section 3. The problem in a nutshell is that although $\Diamond(A \vee B)$ is entailed by, say, $\Diamond A$, and with this the basic free choice reading can be derived, it is not the case that $(A \vee B) > C$ is entailed by $A > C$. This precludes SDA to be derived by iterated innocent exclusion.

Making things more concrete, reconsider the target sentence (18), which we take to implicate that both (19a) and (19b) are true.

(18) If you eat an apple or a pear, you will feel better. $(A \vee B) \boxRightarrow C$

(19a) If you eat an apple, you will feel better. $A \boxRightarrow C$

(19b) If you eat a pear, you will feel better. $B \boxRightarrow C$

For simplicity, it suffices to consider the set of alternatives (the argument also holds if we include the conjunctive alternative):

(56) $\text{ALT} = \{A \boxRightarrow C,\ B \boxRightarrow C,\ (A \vee B) \boxRightarrow C\}$.

In a first round of applying the exhaustivity operator EXH_{IE} to our target sentence we cannot exclude any alternatives innocently, because if $(A \vee B) \boxRightarrow C$ is true, then at least one of $A \boxRightarrow C$ or $B \boxRightarrow C$ must be true as well:

(57) $\text{EXH}_1((A \vee B) \boxRightarrow C) = (A \vee B) \boxRightarrow C$.

This is as with the basic free choice inference we calculated before. But we now get a different prediction for the other alternatives, because, the maximal sets of consistently excludable alternatives to $A \boxRightarrow C$ and $B \boxRightarrow C$ both contain the target sentence:

(58) $\text{EXH}_1(A \boxRightarrow C) = A \boxRightarrow C \wedge \neg(B \boxRightarrow C) \wedge \neg((A \vee B) \boxRightarrow C)$

$\text{EXH}_1(B \boxRightarrow C) = B \boxRightarrow C \wedge \neg(A \boxRightarrow C) \wedge \neg((A \vee B) \boxRightarrow C)$.

But that means that, although at the next application of exhaustification of $(A \vee B) \boxRightarrow C$ both $\textsc{Exh}_1(A \boxRightarrow C)$ and $\textsc{Exh}_1(B \boxRightarrow C)$ can be consistently negated, this will have no effect:

(59) $\textsc{Exh}_2((A \vee B) \boxRightarrow C)$

$= (A \vee B) \boxRightarrow C \wedge \neg\textsc{Exh}_1(A \boxRightarrow C) \wedge \neg\textsc{Exh}_1(B \boxRightarrow C)$

$= (A \vee B) \boxRightarrow C$

$\wedge \neg (A \boxRightarrow C \wedge \neg(B \boxRightarrow C) \wedge \neg((A \vee B) \boxRightarrow C))$

$\wedge \neg(B \boxRightarrow C \wedge \neg(A \boxRightarrow C) \wedge \neg((A \vee B) \boxRightarrow C))$

$= (A \vee B) \boxRightarrow C\,.$

The procedure has reached a fixed point — as the interested reader will happily verify — without enriching the target sentence to support SDA.

In conclusion, the exhaustive-interpretation approach to quantity implicatures seems promising, but it does not yield a uniform explanation for choice-readings of disjunctions and SDA off-the-shelf.[15] This raises the demand for a general account that ideally (i) derives transparently from the assumption that interlocutors behave rationally towards a common shared goal of successful communication, (ii) accounts for the free choice-readings of disjunctions, SDA, as well as epistemic implicatures, in a uniform way, and that also (iii) sheds light on the precise nature of (iterated) exhaustive interpretation. This is what the following game theoretic model does. In order to motivate its general set-up, I would first like to briefly discuss what is normally considered a rational explanation of behavior in general and of conversational implicatures in particular.

5 Rationalizing quantity implicatures

If Grice's conjecture about a possible rational foundation of implicatures is correct, then a quantity implicature is really an *abductive inference* that rationalizes why the speaker has made a certain utterance (cf., Geurts 2010).

15 A reviewer points out correctly that SDA can be derived after all if we assume that $(A \vee B) > C$ is *not* an alternative to $A > C$ and $B > C$. An asymmetric notion of alternativeness is not implausible at all, and has been shown to be capable of interesting explanatory work (cf., Spector 2007). In fact, asymmetry in the set of alternatives would also greatly benefit the system that this paper introduces, in particular with problems of "scaling-up" depicted in Section 9.2 (see also Footnote 32). I eschew following this path presently — for the sake of Fox's account or my own — for want of a better general understanding of the notion of alternativeness.

In simple cases, this inference is still rather perspicuous. It is a piece of hearer reasoning that starts, or so I propose, from the following premises:

(P 1) The speaker uttered φ.

(P 2) The speaker could have uttered ψ also and was aware of that.

(P 3) The speaker rationally chose φ over ψ for a reason.

What we would like to conclude from this is at least the general epistemic implicature:

(C) The speaker does not believe that ψ is true.

which could, of course, be strengthened by competence reasoning where necessary. In what sense and under which circumstances does this conclusion follow from these premises?

The pivotal element clearly is premise 3, the speaker's rationality. A choice of φ over ψ is rational if the speaker believes that φ suits her purpose no worse than a choice of ψ. But that means that a choice can be rational for myriad reasons. With some vivid fantasy we may concoct ever new pairs of beliefs and preferences to ascribe to the speaker, all of which could make her choice a rational one.

This is where the abductive nature of the inference surfaces most clearly: this is not a deduction, but an inference to the best explanation. What a "best explanation" is in a given situation, is not a matter of logical necessity but of common sense. What we are looking for, then, is a set of plausible extra assumptions X about the speaker's mental state, her beliefs and preferences, that satisfies two conditions: (i) the general epistemic implicature should be contained in X, and (ii) the conjunction of X should (defeasibly and non-trivially) entail the speaker's rationality. For instance, the general epistemic implicature in (C) entails the speaker's rationality in conjunction with the following two assumptions about speaker beliefs (A 1) and preferences (A 2):

(A 1) The speaker believes that the hearer will come to believe that ψ is true if she utters it.

(A 2) The speaker wants the hearer to believe a proposition only if she believes it too.

The problem for a systematic grounding of quantity implicatures in terms of rationality is now clear: we would need more clarity concerning additional

assumptions about the speaker's mental state; ideally, we would like to have a perspicuous model of interactive speaker and hearer beliefs and preferences that is derived from a handful of innocuous and plausible assumptions about how interlocutors may construct a representation of the context of utterance when needed, and also of the beliefs that interlocutors have about each others' behavior and beliefs. This is why we should turn to *game theory*, which gives us exactly what we need: (i) a sufficiently detailed and principled context-model, (ii) a formal notion of rationality in terms of agents' beliefs and preferences, and (iii) a systematic way of assessing players' interactive beliefs about beliefs and action choices.[16]

6 Interpretation games

An utterance and its uptake can be represented in terms of a *signaling game*. Signaling games have been studied extensively in philosophy (Lewis 1969), linguistics (e.g., Parikh 2001, van Rooij 2004, Jäger 2007), biology (e.g., Grafen 1990) and economics (e.g., Spence 1973). A signaling game is a simple dynamic game with imperfect information between two players: a sender and a receiver. The sender knows the actual state of the world t, but the receiver doesn't. The sender chooses a message m from a given set of alternatives, all of which we assume here to have a semantic meaning commonly known between players. The receiver observes the sent message m and chooses an action a. An *outcome* of playing a signaling game for one round is given by the triple t, m and a. Each player has his own preferences over such outcomes.

Formally, a signaling game (with meaningful signals) is a tuple

(60) $\langle \{S, R\}, T, \Pr, M, [\![\cdot]\!], A, \mathrm{U}_S, \mathrm{U}_R \rangle$

where sender S and receiver R are the players of the game; T is a set of states of the world; $\Pr \in \Delta(T)$ is a prior probability distribution over T, which represents the receiver's uncertainty which state in T is actual;[17] M is a set of messages that the sender can send; $[\![\cdot]\!] : M \to \mathcal{P}(T) \setminus \emptyset$ is a denotation function that gives the predefined semantic meaning of a message as the

16 For general introduction to game theory in the context of linguistic pragmatics, see Benz, Jäger & van Rooij (2006).

17 As for notation, $\Delta(X)$ is the set of all probability distributions over set X, Y^X is the set of all functions from X to Y, $X\colon Y \to Z$ is an alternative notation for $X \in Z^Y$, and $\mathcal{P}(X)$ is the power set of X.

set of all states where that message is true; A is the set of response actions available to the receiver; and $\mathrm{U}_{S,R} : T \times M \times A \to \mathbb{R}$ are utility functions for both sender and receiver, mapping each outcome $\langle t, m, a \rangle$ to a numerical payoff that represents how desirable this outcome is to the player.

For example, a (trivial) signaling game for the interpretation of an utterance of "Dada is drunk", could contain states t_{drunk} and t_{sober}, in which Dada is drunk or sober respectively. There could be two messages representing the utterances "Dada is drunk" and "Dada is sober". The receiver would respond with an action after he receives a message, such as to buy Dada another schnapps or rather shove him into a taxi, depending on his preferences over outcomes and his beliefs about prior probabilities of states, sender behavior and so on.

Towards an explanation of quantity implicatures in natural language, a special class of signaling games is of particular relevance: those which implement the basic Gricean assumptions of cooperativity and relevance of information. I will refer to signaling games that implement these assumptions as *interpretation games*. Interpretation games are representations of a context of utterance of a to-be-interpreted sentence that capture all and only the essential features of an utterance context that are standardly assumed to back up quantity reasoning. I suggest that these context representations are constructed generically from the usual set of alternatives to the to-be-interpreted expression, together with their logical semantics.

I will make a distinction between *base-level* and *epistemic* interpretation games. The former serve as representations of an utterance context of a hearer who is not consciously taking the speaker's epistemic states into account.[18] Base-level interpretation games are where we derive base-level quantity implicatures. Epistemic interpretation games *do* explicitly accommodate the speaker's epistemic states and it is these context models that we consult to explain epistemic quantity implicatures.

18 I propose that this is natural and happens a lot: when a trusted source — think: your mother — says "I am proud of you" you often *directly* integrate the information "mother is proud of me" into your stock of beliefs without necessarily reasoning *explicitly* about your mother's beliefs and competence on the issue at hand. In other words, it seems to me that often the epistemic dimension of a talk-exchange will not be *consciously represented* at all, unless forced by the context or otherwise necessary. For clarity, this means that I will not generally consider base-level implicatures as derived from epistemic implicatures via competence here.

6.1 Base-level interpretation games

To construct a base-level interpretation game for the interpretation of a sentence S with alternatives ALT, we equate first the set ALT with the set of speaker available messages M. Next, we need to define a reasonable set of state distinctions T that are relevant for quantity reasoning. These distinctions hinge on the available alternatives. Clearly, not every possible way the world could be can be distinguished with any set M, and we should therefore restrict ourselves to only those states of affairs that can feasibly be expressed with the linguistic means at hand. Which are these exactly?

Given a set of propositions ALT and two possible worlds w and v, we say that w and v are ALT-indistinguishable, $w \sim_{\text{ALT}} v$, iff for all $A \in \text{ALT}$ we have $w \in A \Leftrightarrow v \in A$. Quantity reasoning for the interpretation of S based on ALT should look at all those worlds in which S is true, but we may safely lump together worlds that are ALT-indistinguishable. Consequently, the set of base-level state distinctions is:

(61) $\quad T_{\text{BL}} = \{\{w \in S \mid w \sim_{\text{ALT}} v\} \mid v \in S\}$.

Of course, the choice of semantic denotation function $[\![\cdot]\!]$ in our game model is then obvious. A message m_{A} is true in t, $t \in [\![m_{\text{A}}]\!]$, if $t \subseteq A$ for that alternative $A \in \text{ALT}$ that m_{A} represents in our context model.

As for the prior probabilities $\Pr(\cdot)$, since we are dealing with general models of utterance interpretation, we would often not assume that the receiver has biased beliefs about which specific state obtains that are relevant for quantity reasoning.[19] In the absence of interpretation-relevant beliefs, we may make a simplifying assumption that $\Pr(\cdot)$ is a *flat probability distribution*:[20]

(62) $\quad \Pr(t) = \Pr(t') \quad$ for all $t, t' \in T$.

Next, the set of receiver actions is equated with the set of states $A = T$ and the receiver's utilities model his interest in getting to know the true state

19 See Allott (2006) and Franke (2009: §3) for discussion of what prior probabilities in an interpretation game could and could not represent.

20 This last assumption may seem contentious. It is not strictly speaking necessary, but it will simplify the solution concept that this paper introduces dramatically, so much so that we can altogether ignore the precise values of probabilities, which, I believe, is a great relief to the working linguist who may prefer not to have to deal with the details of information theory and probability reasoning. (But see also Geurts & Pouscoulous (2009a) for arguments that beliefs about likely worldly states of affairs often do not seem to impact quantity reasoning.)

of affairs, i.e., getting the right *interpretation* of the observed message:

$$(63) \quad \mathrm{U}_R(t, m, a) = \begin{cases} 1 & \text{if } t = a \\ 0 & \text{otherwise.} \end{cases}$$

The assumption underlying this construction is that quantity reasoning is about coordinating which meaning enrichment of the target message is reasonable given a set of possible meaning distinctions induced by the alternatives. Let me briefly enlarge on this.

Any signaling game embeds the structure $\langle T, \Pr, A, \mathrm{U}_R \rangle$ which is a classical *decision problem* of the receiver. Following van Rooij (2003), we may look at such a decision problem as a generalization of the notion of a question. In the present case this gives a flexible and powerful representation of a *question under discussion*: the receiver's decision problem pins down which state distinctions matter in which way to the (practical) decision of the hearer. In other words, this part of the structure implements what is *relevant* to the hearer. In general, many things could be relevant to a conversation in some sense or another. Some of these senses could easily be implemented in the decision problem contained in a signaling game. As for interpretation games, the particular assumption that these structures implement is that the hearer is interested in quantity reasoning, i.e., that he is interested in drawing any finer meaning distinctions that alternative messages could have made.

This settles the utilities of the receiver. As for the speaker, we should assume that conversation is a *cooperative* effort — at least on the level of such generic context models that back up quantity reasoning. This is easily implemented by defining the sender's utilities in terms of the receiver's as follows:

$$(64) \quad \mathrm{U}_S(t, m, a) = \mathrm{U}_R(t, m, a)\,.$$

Here is a simple example to illustrate this construction. Suppose we are interested in explaining the inference from (4) via (5) to (6d).

(4) Some of Kiki's friends are metalheads.
(5) All of Kiki's friends are metalheads.
(6d) It's not the case that *all* of Kiki's friends are metalheads.

We consider two alternative forms $M = \{m_{\text{some}}, m_{\text{all}}\}$ where, obviously, choice of m_{some} represents an utterance of the sentence (4), and m_{all} corresponds to (5). This allows us to distinguish two states within the denotation of

	$\Pr(t)$	$t_{\exists\neg\forall}$	$t_{\forall}$	m_{some}	m_{all}
$t_{\exists\neg\forall}$	$1/2$	1,1	0,0	1	0
$t_{\forall}$	$1/2$	0,0	1,1	1	1

Figure 4 Interpretation game for "some A's are B's"

	$\Pr(t)$	t_{A}	t_{B}	t_{AB}	$m_{\Diamond\text{A}}$	$m_{\Diamond\text{B}}$	$m_{\Diamond(\text{A}\vee\text{B})}$
t_{A}	$1/3$	1,1	0,0	0,0	1	0	1
t_{B}	$1/3$	0,0	1,1	0,0	0	1	1
t_{AB}	$1/3$	0,0	0,0	1,1	1	1	1

Figure 5 Interpretation game for "$\Diamond(A \vee B)$"

the target message m_{some}: in state $t_{\exists\neg\forall}$ m_{some} is true, while m_{all} is false; in state $t_{\forall}$ both messages are true. Together we obtain the interpretation game in Figure 4. (The table lists the prior probabilities for each state, the payoffs for each state-interpretation pair for S and R respectively, and the semantic meaning of messages with a 1 for truth and a 0 for falsity. Most of this information is redundant when it is clear that we are dealing with interpretation games.) A slightly more complex example is the context model in Figure 5 which is constructed for the interpretation of a sentence like (12a), with alternatives as in (17a) and (17b).

6.2 Epistemic interpretation games

Signaling games standardly incorporate the assumption that the sender knows the actual state. For base-level interpretation games, this means that these context models have a strong speaker competence assumption already built in: the interpretation behavior we derive in these context models serves to account for base-level implicatures unmitigated by considerations of the sender's beliefs and opinions.

In order to extend quantity reasoning to epistemic implicatures we also need to accommodate for the possibility of epistemic uncertainty of the sender. The most conservative way of doing so is to stick to the signaling game framework and to simply "epistemicize" the notion of a state: epistemic interpretation games feature a set of states that distinguishes different epistemic states of the speaker, who is then still perfectly knowledgable about

her own state of mind. The hearer's response actions are interpretations as to which epistemic state the speaker is in. So, which epistemic states of the speaker should the hearer distinguish?

The idea is exactly the same as for base-level games. Epistemic states that are relevant to quantity reasoning about the meaning of S given ALT are obtained from looking at all (non-nontrivial) epistemic states in which the target sentence S is *believed to be true*. In that set, we lump together all those epistemic states which cannot be distinguished by different *belief-value*s the speaker may have with regard to different elements in ALT. Recall that in the classical framework that we work in (see Footnote 4), there are three belief values for any given proposition: (i) the agent believes the proposition is true (belief value "1"), (ii) the agent believes it is false (belief value "0"), or (iii) the agent is uncertain about it (belief value "u"). Let's therefore say that two epistemic states X and Y — sets of possible worlds — are ALT-indistinguishable, $X \sim_{\text{ALT}} Y$, iff for all $A \in$ ALT:

(65) $X \subseteq A \Leftrightarrow Y \subseteq A$ and $X \cap A \neq \emptyset \Leftrightarrow Y \cap A \neq \emptyset$.

The set of state distinctions for an epistemic interpretation game for S given ALT is then:

(66) $T_{\text{E}} = \{\{X \subseteq S \mid X \sim_{\text{ALT}} Y\} \mid Y \subseteq S \wedge Y \neq \emptyset\}$.

Accordingly, the semantic denotation function in epistemic interpretation games should be interpreted as: $t \in [\![m_{\text{A}}]\!]$ iff $\bigcup t \subseteq A$ for the unique alternative A that corresponds to m_{A}. The rest stays, more or less, the same.

Here is a simple example. Suppose we are interested in the epistemic quantity implicatures that may arise from contrasting (4) and (5). We then look at the set of epistemic states where the sender believes that (4) is true that differ according to the sender's beliefs about (5). There are three such states as given in the table in Figure 6. (Where feasible, the indices of states are vectors of truth- or belief-values of messages in the order fixed by the table.)

Epistemic interpretation games can also incorporate various assumptions of the hearer about the speaker's competence. A plausible locus for implementing these assumptions are the prior probabilities. For instance, suppose that priors in the game from above are flat: $a = b$. This would encode the hearer's total uncertainty as to whether the speaker is in one epistemic state or another. On the other hand, if we wanted to implement the hearer's

	$\Pr(t)$	$t_{[1,o]}$	$t_{[1,1]}$	$t_{[1,u]}$	m_{some}	m_{all}
$t_{[1,o]}$	a	1,1	o,o	o,o	1	o
$t_{[1,1]}$	a	o,o	1,1	o,o	1	1
$t_{[1,u]}$	b	o,o	o,o	1,1	1	u

Figure 6 Epistemic interpretation game for contrast "some" vs. "all"

assumption that the speaker is competent in the model in Figure 6, we could do so by assuming that $a > b$. This would implement the idea that the hearer considers it *less likely* that the speaker is uncertain, but that he has otherwise no interpretation-relevant beliefs about the speaker's beliefs.

This can be generalized as follows. If the hearer assumes the speaker to be competent, then states with less uncertainty are considered *a priori* more likely than states with more uncertainty. More concretely still, the assumption of sender competence takes the following form in the present framework:

(67) *Competence Assumption:* If the speaker is (believed) competent, then the prior probability $\Pr(t)$ of information state t is given by a strictly *decreasing* function of the number of alternatives that t is undecided about, i.e., assigns the belief value "u".

Of course, there is a third option, as mentioned in Section 2. The hearer may also assume that the speaker is *not* competent. In that case, we may adopt the following:

(68) *Incompetence Assumption:* If the speaker is (believed) incompetent, then the prior probability $\Pr(t)$ of information states t are given by a strictly *increasing* function of the number of alternatives that t is undecided about, i.e., assigns the belief value "u".

Either assumption still leaves some wiggle room, as to how much difference we allow in probability between states where the speaker is differently competent. In the following, I will assume that probability differences are sufficiently small, because this facilitates calculating the solutions of a game tremendously (see Section 8 and also Appendix B).

7 Strategies, equilibria and explanations

Games like those in Figures 4, 5, and 6 fix certain parameters of the utterance context — arguably all and only those that are relevant to quantity reasoning — and as such put certain constraints on what players of this game could possibly believe about each other. But the game model does not fully determine the players' beliefs and action choices either. Still, we saw in Section 5 that it is important for a rationalization of quantity implicatures to give a detailed specification of what, for instance, the speaker believes the hearer's interpretation of a given sentence will be. Towards this end, we should start by fixing a notion of a player's behavior in a game.

Behavior. A player's behavior in a game is captured in the concept of a *strategy*. A *pure sender strategy* is a function $s \in M^T$ and a *pure receiver strategy* is a function $r \in A^M$. Pure strategies define how a player behaves in each possible information state that she might find herself in during the game. As the sender knows the actual state, she can choose a message conditional on the state she is in. As the receiver does not know the actual state, but only the sent message, he can condition his choice of action only on the message that he observed. We say that a pair $\langle s, r \rangle$ of pure sender and receiver strategies is a *pure strategy profile*.

For example, the game in Figure 4 allows for a number of pure strategies. Obviously, the "empirically correct" play, so to speak, would be:

$$(69) \qquad s = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\text{some}} \\ t_{\forall} & \mapsto & m_{\text{all}} \end{array} \right\} \qquad\qquad r = \left\{ \begin{array}{lcl} m_{\text{some}} & \mapsto & t_{\exists\neg\forall} \\ m_{\text{all}} & \mapsto & t_{\forall} \end{array} \right\}$$

whose receiver part captures drawing the attested quantity inference that the use of m_{some} conveys that the actual state is $t_{\exists\neg\forall}$, and whose sender part also conforms to our intuition about speakers who adhere to the Speaker Quantity Principle in (3). This is not the only pure strategy profile in this game, but this is the one that we would like to be selected — preferably uniquely — by a suitable *solution concept*, that specifies what counts as good or optimal behavior, so as to yield an explanation of a quantity implicature.

Similarly, for the interpretation game in Figure 5, we could say that we had "explained" the attested quantity implicature if by some independently

motivated criterion we could select the pure strategy profile:

$$(70)\qquad s = \left\{ \begin{array}{lcl} t_{\mathrm{A}} & \mapsto & m_{\Diamond \mathrm{A}} \\ t_{\mathrm{B}} & \mapsto & m_{\Diamond \mathrm{B}} \\ t_{\mathrm{AB}} & \mapsto & m_{\Diamond (\mathrm{A} \vee \mathrm{B})} \end{array} \right\} \qquad r = \left\{ \begin{array}{lcl} m_{\Diamond \mathrm{A}} & \mapsto & t_{\mathrm{A}} \\ m_{\Diamond \mathrm{B}} & \mapsto & t_{\mathrm{B}} \\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} & \mapsto & t_{\mathrm{AB}} \end{array} \right\}.$$

Nash equilibrium. The most widely-known solution concept for games is that of a *Nash equilibrium*. The idea behind this notion is that an equilibrium characterizes a *steady state*, i.e., a strategy profile in which no player would strictly benefit if he deviated from that profile given that everybody else would conform. For our signaling games, a pure strategy profile $\langle s, r \rangle$ is a pure Nash equilibrium whenever for all $t \in T$ we have:[21]

(i) $\mathrm{U}_S(t, s(t), r(s(t))) \geq \mathrm{U}_S(t, s'(t), r(s'(t)))$ for all $s' \in M^T$, and

(ii) $\mathrm{U}_R(t, s(t), r(s(t))) \geq \mathrm{U}_R(t, s(t), r'(s(t)))$ for all $r' \in A^M$.

Both of the intuitive strategy profiles in (69) and (70) are Nash equilibria, as is easy to check. But, unfortunately, they are not uniquely so. For instance, a profile which is like (69) but which reverses messages:

$$(71)\qquad s = \left\{ \begin{array}{lcl} t_{\exists \neg \forall} & \mapsto & m_{\mathrm{all}} \\ t_{\forall} & \mapsto & m_{\mathrm{some}} \end{array} \right\} \qquad r = \left\{ \begin{array}{lcl} m_{\mathrm{some}} & \mapsto & t_{\forall} \\ m_{\mathrm{all}} & \mapsto & t_{\exists \neg \forall} \end{array} \right\}$$

is also a Nash equilibrium of the game in Figure 4, in which, moreover, both players are just as well off as in the intuitive equilibrium in (69). To rule out this equilibrium, we could therefore rule out that messages may be used untruthfully (after all, message m_{all} is not true in state $t_{\exists \neg \forall}$).

But even that will not select the desired profiles uniquely for all cases. For consider the following Nash equilibrium of the game in Figure 5:

$$(72)\qquad s = \left\{ \begin{array}{lcl} t_{\mathrm{A}} & \mapsto & m_{\Diamond (\mathrm{A} \vee \mathrm{B})} \\ t_{\mathrm{B}} & \mapsto & m_{\Diamond \mathrm{B}} \\ t_{\mathrm{AB}} & \mapsto & m_{\Diamond \mathrm{A}} \end{array} \right\} \qquad r = \left\{ \begin{array}{lcl} m_{\Diamond \mathrm{A}} & \mapsto & t_{\mathrm{AB}} \\ m_{\Diamond \mathrm{B}} & \mapsto & t_{\mathrm{B}} \\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} & \mapsto & t_{\mathrm{A}} \end{array} \right\}.$$

This is at odds with the intuitive interpretation of the corresponding natural language expressions, but it is nonetheless a proper Nash equilibrium in

21 For notational clarity, $s(t)$ is the message that the sender sends when following strategy s, and $r(s(t))$ is a receiver's response action to that message.

which players achieve perfect communication at maximal payoffs throughout. This example is particularly worrying because there seems to be no obvious refinement of the equilibrium notion that rules out (72) in favor of (70). For instance, Parikh (2001) proposes to filter Nash equilibria by a secondary criterion of *Pareto-optimality*: a strategy profile is Pareto-optimal if any deviation to some player's benefit would be to some other player's detriment. The Nash equilibrium in (72) *is* Pareto-optimal in this sense. Another possible set of refinements is suggested by van Rooij (2008), namely the *Neo-Logism Proofness* criterion of Farrell (1993) and the *intuitive criterion* of Cho & Kreps (1987). Neither of these refinements has any bite on the equilibrium in (72), because there are no "surprise messages" (see below) to which both criteria could apply.

The source of the problem is, arguably, an improper treatment of conventional semantic meaning in pragmatic reasoning (cf., Franke 2009, 2010). Even if we were to assume that speakers cannot speak untruthfully, or are severely punished when they do, traditional solutions are still too weak.[22] What we need is a general solution concept that both properly accounts for the role of semantic meaning and selects uniformly for the "empirically correct" solutions. This is what the solution concept does that the following section will introduce.

8 Iterated best response reasoning

This section introduces a simplified and accessible version of *iterated best response* (IBR) reasoning, as a solution concept that captures pragmatic reasoning on top of a given semantics. A more precise formulation of this solution concept utilizes the standard definitions of probabilistic beliefs, Bayesian updates, expected utilities and the like and is given in Appendix B. This section gives a "light system" that is easier to handle. The light system is derived from the "heavy system" of Appendix B by the extra assumption that priors are (nearly) flat and beliefs about opponent behavior are abstracted from too. Generally speaking, IBR reasoning may be seen as a means of selecting a suitable equilibrium, by factoring in semantic meaning as a fo-

22 More strongly even, it would be methodologically short-sighted, if not plainly wrong to make either assumption, because, as is established wisdom in game theory (Rabin 1990, Farrell 1993, Farrell & Rabin 1996), which aspects of conventional meaning are *credible* is subject to strategic considerations and depends crucially on the degree of conflict of interests between speakers and hearers.

cal starting point *before* rationalizing utterances and interpretations. To appreciate this set-up, a little background is helpful.

Step-wise reasoning & behavioral economics. Recent years have witnessed a massive increase in experimental approaches to strategic reasoning (cf., Camerer 2003). Experimental results strongly suggest that equilibrium solution concepts are, though theoretically appealing, *not* the best predictors of human performance. Rather, there is ample empirical evidence for what has been called "level-k reasoning", i.e., best response reasoning over discrete steps to only a certain depth k (e.g., Ho, Camerer & Weigelt 1998, Crawford 2007, Crawford & Iriberri 2007). Intuitively speaking, such "level-k reasoning" proceeds from some psychologically salient strategy, then first considers a best response to that initially salient strategy, then a best response to that, and so on. When playing a game for the first time, most human players apply 1 or 2 rounds of such iterated reasoning. But on repeated trials agents can learn to apply higher levels of best response reasoning as well (cf., Camerer 2003: §5–6).

Such "level-k reasoning" also plausibly solves the problem of equilibrium selection that we encountered in Section 7. The idea is to treat the semantic meaning of expressions as a focal attractor of attention, i.e., as an initially *salient strategy* from which iterated best response reasoning departs. Unsophisticated level-0 players only stick to the semantic content of expressions: level-0 senders arbitrarily say something true, and level-0 receivers interpret a message literally, i.e., as if it was a true observation revealed by nature, not a strategic choice of a communicating agent. On higher levels of sophistication, level-$(k+1)$ players choose a best response to the belief that their opponent is a level-k player. This process may lead to a fixed point and it is this fixed point behavior which explains, I shall demonstrate, the pragmatic enrichments of messages in interpretation games.[23]

23 This does *not* mean that I suggest that every implicature is established by explicitly calculating through such a sequence of level-k thinking. This would be ludicrous. Just as exhaustivity-based approaches are not meant to be — or so I hope — psychologically realistic descriptions of the actual processes of natural language interpretation, so is the model I propose here. The IBR model wants to describe a generalized and idealized pragmatic competence, but it does so in explicit Gricean, rationalistic terms (see also the discussion in Section 10).

IBR reasoning with flat priors. IBR reasoning is such step-by-step reasoning. We will therefore define inductively differently sophisticated *player types*, one player type for each reasoning step. A player type is defined here as a set of pure strategies: those pure strategies that a player of that level of sophistication may be expected to play. These player types are singled out by particular assumptions about the *epistemic states* of players. In particular, the player type definitions here are motivated by three assumptions. Firstly, we make two assumptions about the hierarchy of player types, namely that it is common belief among players that:

BASE: level-0 players are entirely unstrategic; and that

STEP: level-$(k+1)$ players act rationally based on an *unbiased belief* — to be defined below — that their opponent is a level-k player.

Secondly, we also make an assumption regulating the impact of conventional meaning on players' belief formation and choices, the so-called *truth ceteris paribus* (TCP) assumption, that it is common belief among players that:[24]

TCP: everybody will stick to the conventional semantic meaning if otherwise indifferent.

Assumptions BASE and TCP yield level-0 players whose behavior is unconstrained except for truthfulness and literal uptake. So, as inductive base, define R_0 as the set of all pure receiver strategies that interpret a message as true, and S_0 as the set of all pure sender strategies that send a true message:[25]

$$(73) \qquad R_0(m) = [\![m]\!] \qquad\qquad S_0(t) = R_0^{-1}(t)\,.$$

The inductive step needs a little more elaboration. If S_k and R_k are sets of pure sender and receiver strategies respectively, the sets S_{k+1} and R_{k+1} are defined as all *rational* choices given some belief that the opponent plays a strategy in S_k or R_k. This could in principle be implemented in many different ways. Most importantly, we could adopt many *prima facie* plausible

24 This formulation is much stronger than actually needed. It is needed here merely to assure semantics-conform behavior after so-called "surprise messages" and when interpretations are "uninducible" (see in particular Section B.3).

25 As for notation, in the present context it is feasible and helpful to represent a set of pure strategies $X \subseteq Z^Y$ as listing for each $y \in Y$ the set $X(y)$ of all $z \in Z$ such that for some strategy $x \in X$ we have $x(y) = z$. I will then also write $X^{-1}(z)$ for the set $\{y \in Y \mid \exists x \in X \exists z \in Z\colon x(y) = z\}$.

constraints on the belief formation process of level-$(k+1)$ players. To keep the system simple, I will assume here that players form *unbiased beliefs*:[26] a level-$(k+1)$ sender (receiver) believes that any strategy in R_k (S_k) is *equally likely*. Level-$(k+1)$ behavior is defined as rational under this belief.

A more traditional rendering of probabilistic beliefs and rational choice is spelled out in Appendix B.1. But if we assume unbiased beliefs, the relevant reasoning can also be appreciated in more intuitive terms. Take the case of a level-$(k+1)$ sender, who believes — by the STEP-assumption — that if she sends m the receiver will choose any interpretation from $R_k(m) \subseteq T$ with probability $|R_k(m)|^{-1}$. We then need to ask, what is a rational choice, given this belief, if the sender is in state t? Since the sender wants to induce the correct interpretation t, we first look at the set of messages that could in principle trigger interpretation t:

(74) $\quad R_k^{-1}(t) = \{m \in M \mid \exists r \in R_k\colon\ r(m) = t\}\ .$

This set could be empty, and if it is, this means that the sender believes that interpretation t is *not inducible by any message*. She would therefore be indifferent as to which message to send as far as communicative success is concerned, but due to the TCP assumption, she would at least choose any message that is true in t. If, on the other hand $R_k^{-1}(t)$ contains at least one message which could trigger the right interpretation, then a best choice of the sender is any message in $R_k(t)$ which makes it *most likely* that t is chosen. This will be messages in $R_k(t)$ for which the set

(75) $\quad R_k(m) = \{t \in T \mid \exists r \in R_k\colon\ r(m) = t\}$

is minimal. Taken together, a level-$(k+1)$ sender's rational choices in t are given by:

(76) $$S_{k+1}(t) = \begin{cases} \arg\min_{m \in R_k^{-1}(t)} |R_k(m)| & \text{if } R_k^{-1}(t) \neq \emptyset \\ S_0(t) & \text{otherwise.} \end{cases}$$

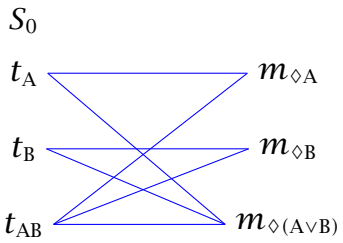
26 Different systems of "level-k reasoning" differ exactly in this design choice. The *cognitive hierarchy model* of Camerer, Ho & Chong (2004), for example, assumes that players of level $k+1$ form the belief that their opponent is of level $l \leq k$. This may seem more realistic, but strongly complicates the mathematics. Other viable options are implemented by Jäger & Ebert (2009) or Mühlenbernd (2009). For our present purposes, the assumption of unbiased beliefs is welcome, chiefly because it helps simplify the system tremendously. But the assumption of unbiased beliefs also does some good pragmatic work for us: it implements a particular form of *forward induction reasoning* without which weaker systems, such as that by Jäger & Ebert (2009), cannot derive, for instance, the free choice implicatures of (12a) (see Franke 2009: §2).

Surprisingly enough, despite the asymmetry in available information between players, the same reasoning also applies to computing sophisticated receiver types if we assume that the prior probabilities are flat (see the proof of Theorem 1 for details). The set of a level-$(k+1)$ receiver's best responses to m is:

$$R_{k+1}(m) = \begin{cases} \arg\min_{t \in S_k^{-1}(m)} |S_k(t)| & \text{if } S_k^{-1}(m) \neq \emptyset \\ R_0(m) & \text{otherwise.} \end{cases} \tag{77}$$

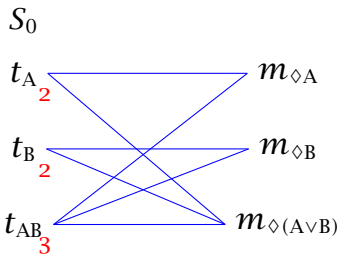
Solving games with diagrams. There is a fairly easy algorithmic way of computing IBR reasoning with the help of simple diagrams. A set of pure strategies, such as for instance S_0 from the game in Figure 5, can be represented by the corresponding mappings of all $S_0(t)$ as follows (the mapping is in left-to-right direction):

(78)



If we want to compute the best response of R_1 to this behavior, we simply need to count the number of outgoing connections from each state:

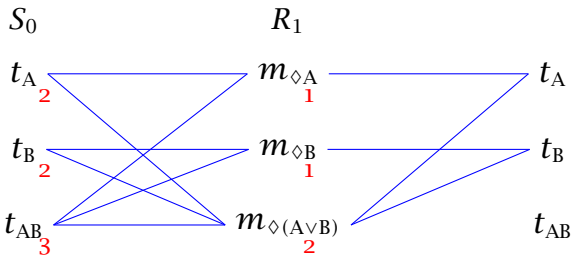
(79)



To plot our level-1 receiver we will then draw a connection from each message m to all those states that are connected with m that have the lowest number

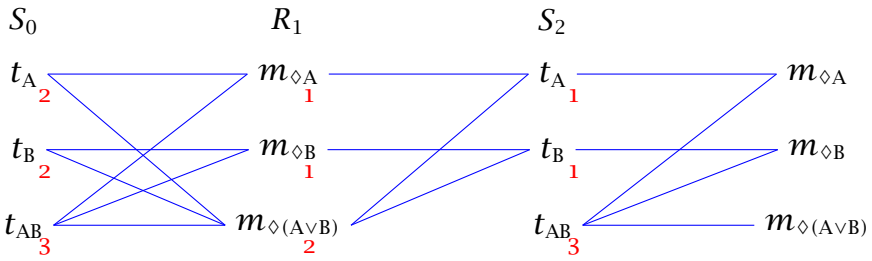
of outgoing connections among states connected with m:

(80)



Whenever, as in the computation of S_2, a node has no incoming connections, we restore the connections from the initial level-0 mapping:

(81)



If we apply the same procedure twice more, we enter into a fixed point in which sender and receiver both play a unique pure strategy, namely:

(82) $$s = \left\{ \begin{array}{lcl} t_A & \mapsto & m_{\Diamond A} \\ t_B & \mapsto & m_{\Diamond B} \\ t_{AB} & \mapsto & m_{\Diamond(A\vee B)} \end{array} \right\} \qquad r = \left\{ \begin{array}{lcl} m_{\Diamond A} & \mapsto & t_A \\ m_{\Diamond B} & \mapsto & t_B \\ m_{\Diamond(A\vee B)} & \mapsto & t_{AB} \end{array} \right\}.$$

This is indeed the strategy profile we want to see selected. But we should also check the predictions of the IBR reasoning chain that starts with an unsophisticated receiver.[27] The full reasoning sequence of both IBR chains is represented in Figure 7. Indeed, both strands of IBR reasoning single out the "empirically correct" behavior. In a sense to be made precise below, this fixed point behavior would then explain the free choice inference as a matter of rational language use.

27 Other related approaches consider only pragmatic reasoning that starts from the assumption of a naïve listener (e.g., Stalnaker 2006, Benz & van Rooij 2007). Since I know of no good reason not to also assume that pragmatic reasoning might depart from the assumption of a naïve speaker too, I would like to stay on the safe side and check predictions of both approaches.

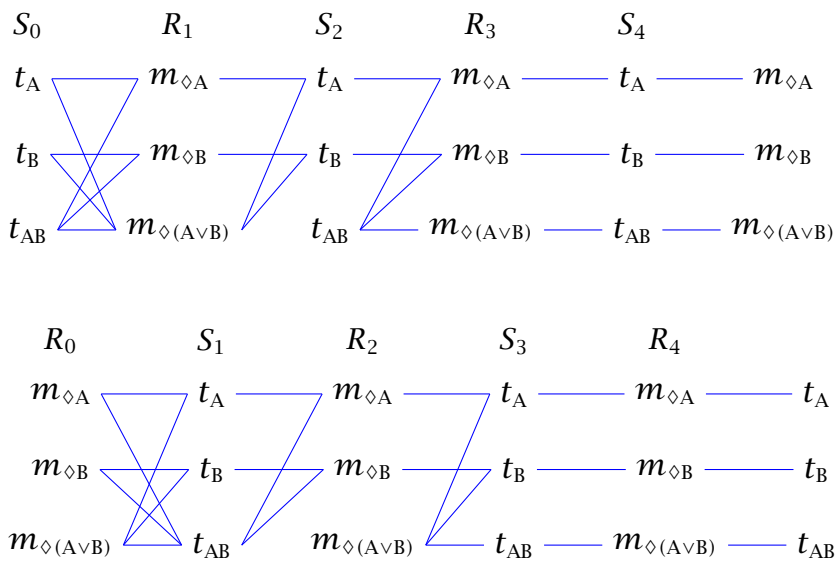


Figure 7 Schematic IBR reasoning for the game in Figure 5

IBR reasoning with non-flat priors. The above definition (77) of rational receiver behavior applies whenever an interpretation game has flat priors. This also covers epistemic interpretation games without competence or incompetence assumption, such as when $a = b$ in the game in Figure 6. The result of running the IBR algorithm in this case is plotted in Figure 8. Both strands of IBR reasoning lead to the general epistemic implicature that the speaker does not believe that the alternative "all" is true.

Unfortunately, we cannot use the same algorithm to solve interpretation games with non-flat priors. Still, this does not necessarily mean that we always have to compute all probabilities and expected utilities in detail. If the differences between the prior probabilities are *small enough* we can treat them *as if* they were a second-order selection criterion on top of the interpretations selected under flat priors. This makes for a similarly easy algorithmic implementation of IBR reasoning in games with non-flat priors.

Theorem 2, given in Appendix B.2, tells us that when differences between priors are small enough we can calculate the best responses of a level-$(k+1)$ receiver as follows:

$$\breve{R}_{k+1}(m) = \begin{cases} \arg\min_{t \in S_k^{-1}(m)} |S_k(t)| & \text{if } S_k^{-1}(m) \neq \emptyset \\ R_0(m) & \text{otherwise} \end{cases} \tag{83}$$

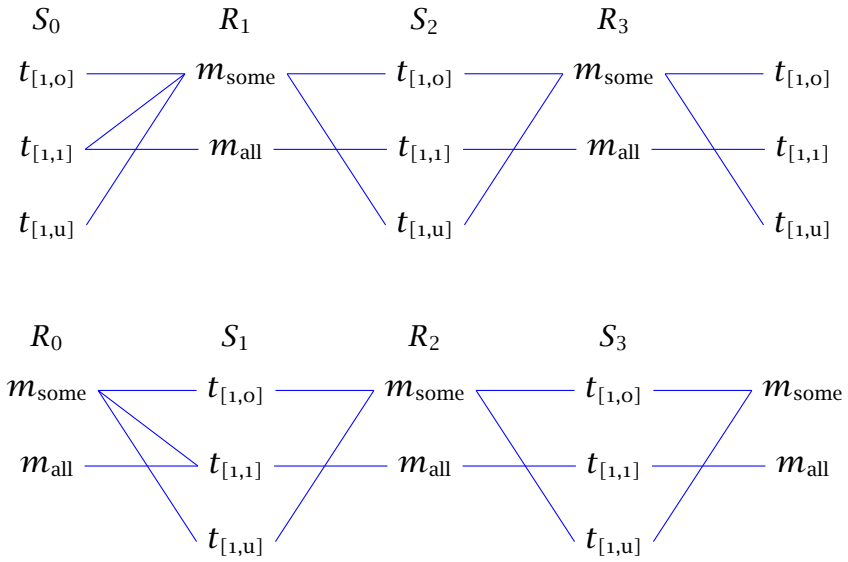


Figure 8 Predictions for the game in Figure 6 when $a = b$

$$R_{k+1}(m) = \left\{ t \in \breve{R}_{k+1}(m) \mid \neg \exists t' \in \breve{R}_{k+1}(m) \colon \Pr(t') > \Pr(t) \right\} .$$

In words, we first compute the receiver's possible interpretations $\breve{R}_{k+1}(m)$ as before and then choose from $\breve{R}_{k+1}(m)$ those states that maximize $\Pr(\cdot)$. That means that if differences in priors are small enough, $\Pr(\cdot)$ applies *as if* this was a secondary selection criterion in a lexicographic ordering after we have evaluated regular quantity reasoning.

For instance, if we look at the game in Figure 6 again, and if we adopt the competence assumption in (67) to the effect that a is *slightly* bigger than b, then we can calculate $R_1(m_{some})$ in three steps (see Figure 9 for the full results). Firstly, we look at all the states where m_{some} gets sent. This is $[\![m_{some}]\!] = \{t_{[1,o]}, t_{[1,1]}, t_{[1,u]}\}$. Next, we select for the states in $[\![m_{some}]\!]$ in which fewest signals are chosen by S_0. These are $t_{[1,o]}$ and $t_{[1,u]}$. Finally, we select the *a priori* most likely states from these. This leaves us with just $t_{[1,o]}$. In the resulting fixed point, the hearer believes that a speaker who produces m_{some} believes that m_{all} is false. Similarly, under the sender incompetence assumption in (68) message m_{some} is associated with $t_{[1,u]}$ in both fixed points. This captures the inference that the speaker has no opinion as to whether the alternative m_{all} was true.

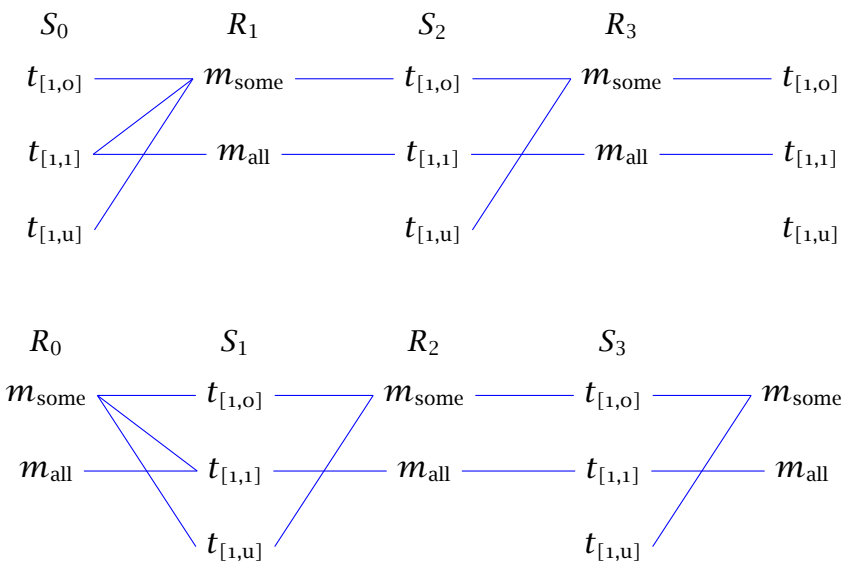


Figure 9 Predictions for the game in Figure 6 when $a > b$

Interpretation of fixed-point behavior. When it comes to matching empirical data, it is the fixed points of IBR reasoning that concern us most, because the behavior selected by any fixed point is compatible with what the most sophisticated agents in our model would play. (In fact, fixed point behavior is rational behavior compatible with common belief in rationality.) In the above examples, the fixed points then, in a way of speaking, *selected* the appropriate Nash equilibria that capture the attested implicatures. If the "correct" equilibrium is selected in this way, this explains the attested implicatures as the outcome of rational language use given a fixed semantics.

In general, IBR reasoning will always reach a fixed point (in finite interpretation games), which will always be a (mild refinement of a) Nash equilibrium.[28] This is the content of Theorems 3 and 4, stated and proved in Appendix B.4. Of course, it remains to be seen whether IBR reasoning always selects the "correct" equilibrium. This is what we turn to next.

28 Notice that it is, however, not necessarily the case that each IBR reasoning sequence reaches the *same* fixed point (see Footnote 29 above). I currently know of no natural condition that would depict the class of games where both IBR reasoning strands necessarily converge. This issue, however, is also only of marginal relevance to pragmatic applications.

9 Predictions & results

This section is dedicated to checking some of the predictions of the IBR model. Section 9.1 deals with the implicatures of the three types of disjunctive constructions that were discussed in Section 3. Section 9.2 considers some further applications and some potential problems of this approach, with indications for possible solutions. The results of this section are summarized in Figure 15 on page 51.

9.1 Disjunctions: Checking the test cases

We would like to check whether the IBR model can account for all of the implicatures standardly associated with: (i) *plain disjunctions* of the form $A \vee B$, (ii) FC-*disjunctions* of the form $\Diamond(A \vee B)$, and (iii) SDA-*disjunctions* of the form $(A \vee B) > C$ (see Figure 2 on page 15). In particular, we'd like to derive two *base-level implicatures*: the free choice inference for $\Diamond(A \vee B)$ and the SDA inference for $(A \vee B) > C$. Moreover, we would like to derive *epistemic implicatures*: a plain disjunction $A \vee B$ is associated with the *ignorance implicature* that the speaker is uncertain about either disjunct: $\mathrm{Uc}_S A \wedge \mathrm{Uc}_S B$; an FC-disjunction can gives rise to $\mathrm{Uc}_S \Diamond A \wedge \mathrm{Uc}_S \Diamond B$; and an SDA-disjunction can give rise to $\mathrm{Uc}_S(A > C) \wedge \mathrm{Uc}_S(B > C)$. Additionally, we will also check whether the IBR model can derive *exclusivity implicatures* as quantity implicatures if we also take into account the respective conjunctive alternatives.

Base-level implicatures. We have already seen how the IBR model deals with the basic free choice readings of sentences of the form $\Diamond(A \vee B)$ as a base-level implicature in a base-level interpretation game. The context model was given in Figure 5 and it had three state distinctions, repeated here:

(84)

	$m_{\Diamond \mathrm{A}}$	$m_{\Diamond \mathrm{B}}$	$m_{\Diamond(\mathrm{A} \vee \mathrm{B})}$
t_{A}	1	0	1
t_{B}	0	1	1
t_{AB}	1	1	1

If we assume that a plain disjunction $A \vee B$ has alternatives A and B, then we derive three isomorphic state distinctions:

(85)

	m_{A}	m_{B}	$m_{\text{A}\vee\text{B}}$
t_{A}	1	0	1
t_{B}	0	1	1
t_{AB}	1	1	1

Interestingly, the same holds for conditionals of the form $(A \vee B) > C$ with alternatives $A > C$ and $B > C$ under the simple order-sensitive semantics given in Section 3:

(86)

	$m_{\text{A}>\text{C}}$	$m_{\text{B}>\text{C}}$	$m_{(\text{A}\vee\text{B})>\text{C}}$
t_{A}	1	0	1
t_{B}	0	1	1
t_{AB}	1	1	1

That means that all three constructions give rise to the same context models, not only at base-level but also for epistemic interpretation games (see below). Consequently, we only have to check predictions of the IBR model once, as these will be identical for all three cases.

Indeed, we have already done so. The calculation for the game from Figure 5 was graphically depicted in Figure 7. The exact same reasoning applies to the other two cases, provided we reinterpret the state names according to the tables above. This result is good and bad. It is good, because free choice and SDA are treated exactly alike by the IBR model. This improves on the predictions of exhaustivity-based approaches. But, on the other hand, IBR also predicts that in this context model a plain disjunction $A \vee B$ is associated with the conjunctive meaning that *both* A and B are true! This is a curious result, and we will come back to this issue in Section 9.2 after we have assessed the predictions for plain disjunctions in other context models as well.

Epistemic implicatures. The epistemic states that we can distinguish based on a plain disjunction $A \vee B$ with alternatives A and B are the following six

states:

(87)

	m_{A}	m_{B}	$m_{\text{A}\vee\text{B}}$
$t_{[1,0,1]}$	1	0	1
$t_{[0,1,1]}$	0	1	1
$t_{[1,1,1]}$	1	1	1
$t_{[1,u,1]}$	1	u	1
$t_{[u,1,1]}$	u	1	1
$t_{[u,u,1]}$	u	u	1

It is clear that six exactly parallel patterns of belief value distributions arise for $\Diamond(A \vee B)$ and $(A \vee B) > C$. It suffices therefore to stick with the case of plain disjunctions.

In order to encode different competence assumptions into the context, we should parameterize the prior probabilities as follows (recall that the number of 'u'-s is relevant in assumptions (67) and (68)):

(88)

	$t_{[1,0,1]}$	$t_{[0,1,1]}$	$t_{[1,1,1]}$	$t_{[1,u,1]}$	$t_{[u,1,1]}$	$t_{[u,u,1]}$
$\Pr(\cdot)$	a	a	a	b	b	c

We'd like to check that for any choice of parameters a, b and c, the proper ignorance reading is derived. This means that we would like to find $t_{[u,u,1]}$ as the only interpretation assigned to $m_{\text{A}\vee\text{B}}$ in all fixed points. In this state, the speaker is uncertain about A and B, but she believes in $A \vee B$. For plain disjunction $A \vee B$, this amounts to the ignorance implicature $\text{Uc}_S A \wedge \text{Uc}_S B$. (For the other disjunctive constructions the same applies with due changes in interpretation of $t_{[u,u,1]}$.)

This is indeed the only interpretation selected for the target message in all six constellations we would need to check. (Two IBR sequences for three different parameter sets.) Here is an informal argument why this is so. For instance, R_1 will interpret $m_{\text{A}\vee\text{B}}$ as $t_{[u,u,1]}$, and only as $t_{[u,u,1]}$, because this is the state which minimizes the number of true messages in this game. For the same reason, S_1 will use $m_{\text{A}\vee\text{B}}$, and only $m_{\text{A}\vee\text{B}}$, in state $t_{[u,u,1]}$. Later steps of either sequence will not change this one-to-one association, and different parameters implementing different competence assumptions do not change this mapping either. (As one example of the six necessary calculations, the sequence starting with S_0 without competence assumption is given in Figure 10; the other calculations are boringly similar.)

Different competence assumptions do however influence the interpretation of messages other than $m_{\text{A}\vee\text{B}}$. In the absence of a competence assump-

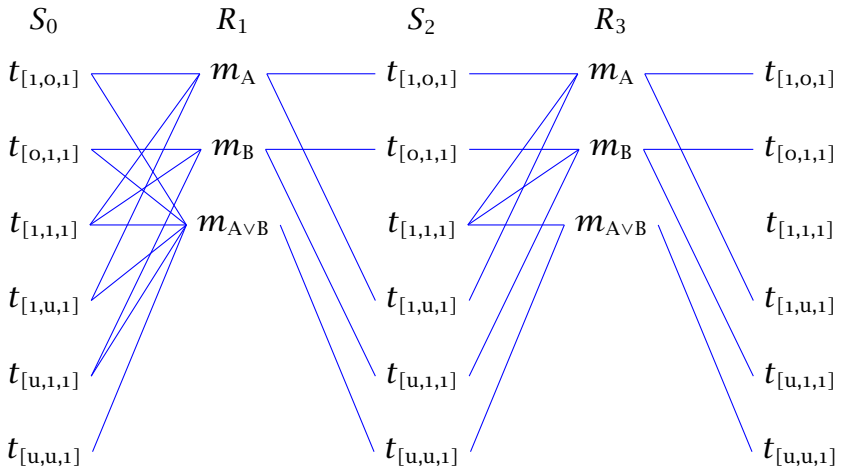


Figure 10 Predictions for epistemic interpretation of "A or B" for $a = b = c$

tion, the unique fixed point of IBR reasoning has m_A interpreted as either $t_{[1,0,1]}$ or as $t_{[1,u,1]}$ (see Figure 10). This captures the reading that the speaker believes that A is true, but does not believe that B is true. On the other hand, if we assume that the speaker is competent, we derive that m_A is interpreted as $t_{[1,0,1]}$: the implicature is that the speaker knows that B is false. Lastly, under an incompetence assumption, we derive the interpretation $t_{[1,u,1]}$ for m_A which captures the reading that the speaker is uncertain about B. All of this accords neatly with intuition.

Exclusivity implicatures at base-level. When we add a conjunctive alternative to the brew, the three types of disjunctive constructions we consider here no longer give rise to the same context models, so that we have to consult each case in turn. The contextual state distinction for base-level interpretation given $A \vee B$ with alternatives A, B and $A \wedge B$ are actually the same as before:

(89)

	m_A	m_B	$m_{A\wedge B}$	$m_{A\vee B}$
t_A	1	0	0	1
t_B	0	1	0	1
t_{AB}	1	1	1	1

Nonetheless, the presence of an additional message will change the interpretation process. The R_0-part of the reasoning is summarized in Figure 11.

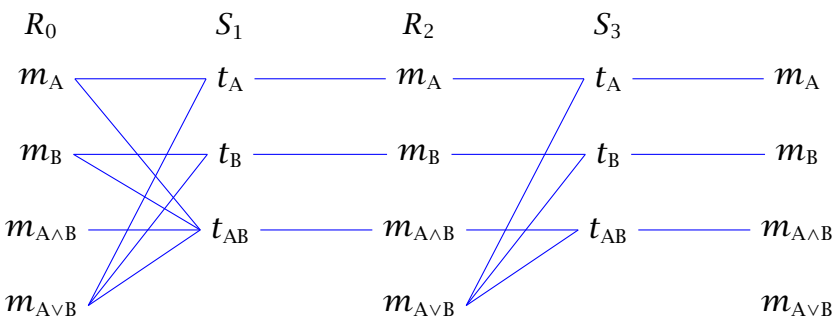


Figure 11 Predictions for plain disjunctions with conjunctive alternative

The S_0-part is analogous and yields an identical fixed point in which, crucially, the disjunction is a so-called *surprise message* to the receiver under a belief in fixed-point sender behavior. A message is a surprise message under a given belief if, according to that belief, the message would never get sent (see also Appendix B.3). Let's make a mental note of this for the subsequent discussion in Section 9.2.

For the target form $\Diamond(A \vee B)$ we derive a different set of state distinctions if we additionally consider the conjunctive alternative $\Diamond(A \wedge B)$, namely:

(90)

	$m_{\Diamond A}$	$m_{\Diamond B}$	$m_{\Diamond(A\wedge B)}$	$m_{\Diamond(A\vee B)}$
$t_{[1,0,0,1]}$	1	0	0	1
$t_{[0,1,0,1]}$	0	1	0	1
$t_{[1,1,1,1]}$	1	1	1	1
$t_{[1,1,0,1]}$	1	1	0	1

Unlike for plain disjunctions, a state $t_{[1,1,0,1]}$ is possible: it is possible for $\Diamond A$ and $\Diamond B$ to be true (there is an accessible A-world, and an accessible B-world), while $\Diamond(A \wedge B)$ is false (there is no accessible world in which both A and B are true). Still, the model's predictions for this case bear no surprises. Figure 12 shows the S_0-sequence. The fixed point interpretation of $\Diamond(A \vee B)$ establishes the free choice reading, as well as the exclusivity implicature that $\Diamond(A \wedge B)$ is false.

Finally, the context model for base-level interpretation of $(A \vee B) > C$ is even a little bigger. We can now distinguish six states, because it is also possible for either of $A > C$ or $B > C$ to be false when $(A \wedge B) > C$ is true.

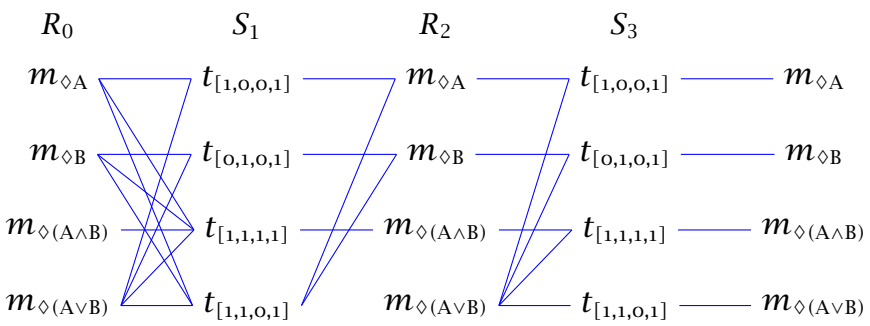


Figure 12 Predictions for FC-disjunctions with conjunctive alternative

So, we get:

(91)

	$m_{A>C}$	$m_{B>C}$	$m_{(A\wedge B)>C}$	$m_{(A\vee B)>C}$
$t_{[1,0,0,1]}$	1	0	0	1
$t_{[0,1,0,1]}$	0	1	0	1
$t_{[1,0,1,1]}$	1	0	1	1
$t_{[0,1,1,1]}$	0	1	1	1
$t_{[1,1,0,1]}$	1	1	0	1
$t_{[1,1,1,1]}$	1	1	1	1

The predictions of the IBR model are summarized in Figure 13, where the R_0 sequence is spelled out. We see that the target message $(A \vee B) > C$ is interpreted correctly in the fixed point that is reached after R_4 to give rise to both the SDA inference, as well as the exclusivity implicature. The results for the S_0-sequence are identical for the interpretation of the target message.[29]

Exclusivity implicatures at epistemic level. It remains to be checked what happens in epistemic interpretation games when we also have the conjunctive alternatives around. Although the context models differ for plain, FC- and SDA-disjunctions, and although calculations are therefore not precisely the same, predictions are essentially the same in all three cases. I will therefore content myself with discussing only the case of plain disjunction, for which

29 The S_0-sequence differs slightly only in the interpretation of $(A \wedge B) > C$.

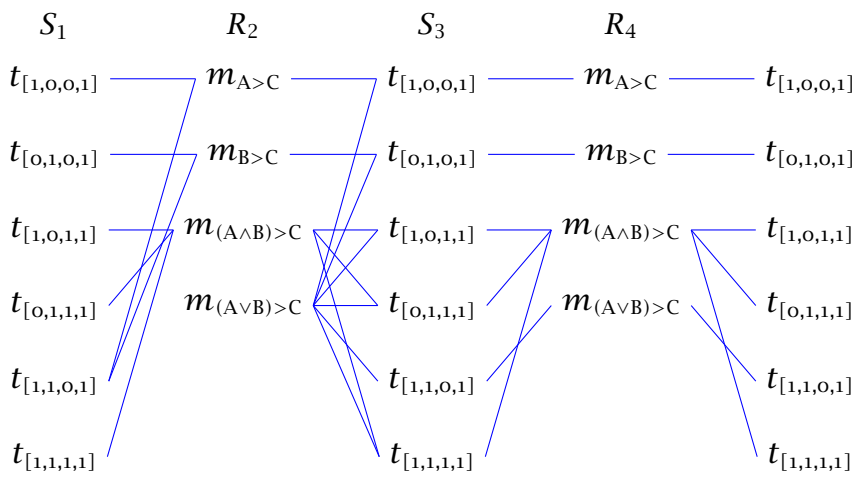


Figure 13 Predictions for SDA-disjunctions with conjunctive alternative

we need to consider the following epistemic state distinctions:

(92)

	m_A	m_B	$m_{A\wedge B}$	$m_{A\vee B}$
$t_{[1,0,0,1]}$	1	0	0	1
$t_{[0,1,0,1]}$	0	1	0	1
$t_{[1,1,1,1]}$	1	1	1	1
$t_{[u,u,0,1]}$	u	u	0	1
$t_{[1,u,u,1]}$	1	u	u	1
$t_{[u,1,u,1]}$	u	1	u	1
$t_{[u,u,u,1]}$	u	u	u	1

An example calculation for the sequence starting with S_0 in the absence of a competence assumption is given in Figure 14. (The other derivations are similar.) In this case we derive the general epistemic inference that the use of $A \vee B$ is associated with the set $\{t_{[u,u,0,1]}, t_{[u,u,u,1]}\}$ which captures that the speaker does not believe that $A \wedge B$ is true. If we integrate the competence assumption, we derive the implicature that $m_{A\vee B}$ is associated with the state $t_{[u,u,0,1]}$ only: the stronger implicature that the speaker knows that $A \wedge B$ is false. Finally, with an incompetence assumption, we derive that the target message is associated with $t_{[u,u,u,1]}$, which vindicates the implicature that the speaker is strictly uncertain about the truth of $A \wedge B$.

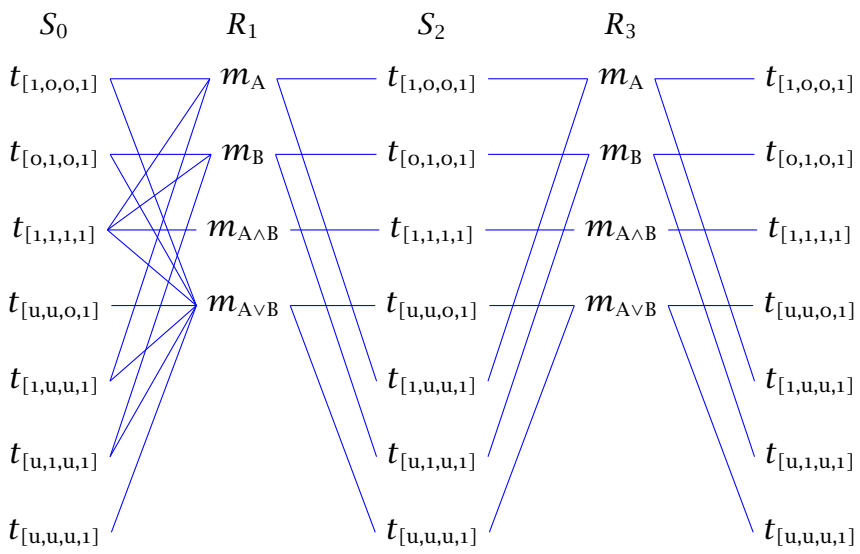


Figure 14 Predictions for the interpretation of "A or B" with conjunctive alternative and no competence assumption

9.2 Reflection: Problems & prospects

So far, the IBR model has done a fair job in explaining the data in accordance with intuition, all of this derived in a formally rigorous account of rational language use and interpretation. The table in Figure 15 provides a summary of the results. But the IBR model is not flawless. The remainder of this section zooms in on some of its problems, together with some thoughts on how these could be overcome. The upshot of the following considerations is that the most fruitful extensions of this paper's proposal concern the context model construction in various ways.

Plain disjunctions at base-level. The results discussed just previously were excellent, except for the interpretation of plain disjunctions in base-level interpretation games. Let's recall what the situation was. If $A\vee B$ is interpreted *without* explicit contrast to the conjunctive alternative $A \wedge B$, then $A \vee B$ is enriched to acquire the meaning of the conjunctive alternative: $A \vee B$ gets to mean $A \wedge B$. If, on the other hand, the conjunctive alternative is explicitly represented in the context model, then $A \vee B$ is not enriched at all, but rather a surprise message that is expected not to be used at all. Finally, the predictions for $A \vee B$ in epistemic game models were fully in accordance with

intuition.

	game model			
	base-level		epistemic	
implicature	w/o conj.	with conj.	w/o conj.	with conj.
plain disjunction				
ignorance	–	–	✓	✓
exclusivity…	–	–	–	✓
…base-level	–	–	–	–
…epistemic	–	–	–	✓
FC disjunction				
free choice	✓*	✓*	–	–
ignorance	–	–	✓	✓
exclusivity…				
…base-level	–	✓*	–	–
…epistemic	–	–	–	✓
SDA disjunction				
SDA	✓*	✓*	–	–
ignorance	–	–	✓	✓
exclusivity…				
…base-level	–	✓*	–	–
…epistemic	–	–	–	✓

Figure 15 Results of the IBR model for the implicatures listed in Figure 2. Checkmarks indicate that an implicature can be derived in a context model type. Asterisks indicate that extra assumptions are necessary for cases with more than two disjuncts.

What are we to make of this? First of all, it needs to be stressed that the problem is *not* that the IBR model makes a wrong prediction. The problem is rather that the use of plain disjunctions that we are after is at odds with the assumptions inherent in base-level interpretation games, most palpably: that the speaker is perfectly knowledgable. The interpretation of plain disjunctions usually requires an epistemic context model.[30] This is to say

30 This intuition is probably also what underlies certain non-standard semantics of disjunctions as inherently modal elements (Zimmermann 2000, Geurts 2005).

that the base-level predictions of IBR reasoning are not wrong as such, but are a result obtained from administering inadequate contextual assumptions. The obvious question to ask next is whether the IBR model helps explaining why plain disjunctions usually want to be interpreted in epistemic context models. Perhaps it does. Here is a try.

If we compare $A \vee B$ directly with $A \wedge B$ at base-level, the former comes out as an unexpected surprise message that requires the hearer to revise his beliefs. One obvious and reasonable way for the receiver to do so is to revise in particular his beliefs about the context model and to adopt an epistemic perspective. And even if $A \wedge B$ is not explicitly compared to $A \vee B$ at base-level, a proficient language user should still realize that the interpretation he may have established for $A \vee B$ could be expressed at equal effort also with other linguistic means such that, if they were taken into account, they would lead to epistemic context models via the above argument. This last step is crucially different for FC- and SDA-disjunctions where the base-level interpretation is *not* exactly that of the respective conjunctive alternative, and where a wide-scope disjunction is clearly more complex.

This explanation is sketchy and leaves many questions unanswered. Still, it seems plausible that proficiency in language also entails proficiency in constructing a proper representation of the context of utterance (where a "proper" context representation is one that proves successful in communication with others). A detailed treatment of this issue is beyond the scope of this paper, but the ideas of Section 6 may hopefully provide a reasonable point of departure for future research.

Entailing disjuncts. So far we have assumed that disjuncts were logically independent. But not all disjunctions have this property, and it's there that we may expect problems with the current approach. For example, under truth-conditional semantics the sentence "A or (A and B)" is equivalent to the sentence "A." But, intuitively, these forms are to be interpreted differently, at least in certain contexts and if we assume that the speaker is competent on the issue at hand. Compare the answers (94a) and (95a) to a question (93).

(93) Who (of John and Mary) came to the party?

(94) a. John did.
b. The speaker knows that John came and that Mary did not.

(95) a. John or (John and Mary).
b. The speaker knows that John came and considers it possible that Mary came too.

Though equivalent in terms of truth-conditions, the implicatures associated with these answers, (94b) and (95b) respectively, are clearly different. This problem palpably affects pretty much all global Neo-Gricean accounts that rely on truth-conditional semantics. Indeed, entailing disjuncts, especially in embedded positions, seem to provide substantial evidence for syntactic/localist approaches to quantity implicatures (cf., Chierchia et al. 2008, 2009, Fox & Spector 2009). It is thus interesting to see whether IBR can cope with this.

Of course, the most reasonable rejoinder here is to deny that truth-conditional semantics is suitable in the first place. The sentences (94a) and (95a) clearly do not have the same dynamic properties as evidenced in (96).

(96) a. ?John came to the party. The latter possibility is rather unlikely though.
b. John or (John and Mary) came to the party. The latter possibility is rather unlikely though.

This is why, for instance, Schulz & van Rooij (2006) use exhaustive interpretation in minimal *dynamic* models to account for cases of this kind. The IBR model could very well do the same thing, as it is not dependent on *any particular* notion of semantics, as long as meaning can be expressed in set-theoretic terms. There is, however, also another possibility worth considering, and that is to assume that the sentences A and $A \vee (A \wedge B)$ are, though equivalent, differently *costly*.

Suppose that for an interpretation of $A \vee (A \wedge B)$ we compare it simply to its disjuncts A and $A \wedge B$. This derives the following set of state distinctions for an epistemic interpretation game:

(97)

	m_{A}	m_{AB}	$m_{\mathrm{A}\vee(\mathrm{A}\wedge\mathrm{B})}$
$t_{[1,0,1]}$	1	0	1
$t_{[1,1,1]}$	1	1	1
$t_{[1,u,1]}$	1	u	1

Let us assume that message $m_{\mathrm{A}\vee(\mathrm{A}\wedge\mathrm{B})}$ is slightly more costly than its equivalent m_{A}. Let us also assume that these costs are *nominal*, as the economists would say, i.e., small enough that they apply — like prior probabilities in our IBR approach — as a *secondary selection criterion*. The IBR reasoning starting with R_0 under a competence assumption is spelled out in Figure 16. The crucial step in this computation is S_1 where message costs favor the sending of m_{A} in states $t_{[1,0,1]}$ and $t_{[1,u,1]}$ over sending $m_{\mathrm{A}\vee(\mathrm{A}\wedge\mathrm{B})}$.

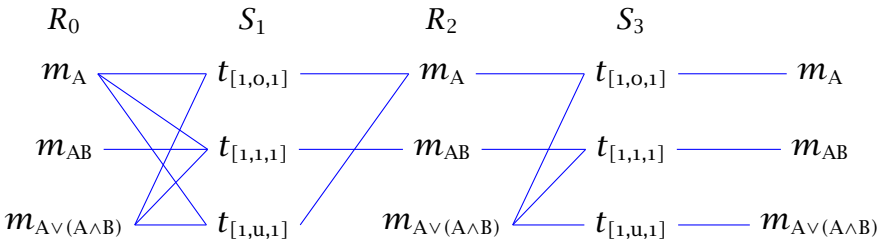


Figure 16 Predictions for interpretation of "A or (A and B)" under a competence assumption

In sum, these considerations point to two interesting extensions of the IBR approach outlined here: firstly, we could consider supplying a non-standard semantics; secondly, we could reflect on the benefits and detriments of exploiting reasoning about message costs. It remains to be seen in how far either of these possibilities would enable the IBR model to account for the intricate empirical data discussed, *inter alia*, by Chierchia et al. (2008, 2009) and Fox & Spector (2009).

Scaling up. Example calculations so far only dealt with disjunctive constructions with two disjuncts. But do we still derive correct predictions when we have more than two? The answer is a hesitant "yes-and-no". It is a "yes-and-no" because some predictions scale up smoothly, others don't (see also Figure 15 for summary). It is hesitant because I actually do not believe that we necessarily have to ask this question in the first place (or, rather: ask it and hope for an affirmative answer). Let me enlarge briefly, on the "yes-and-no", sketching some of the results, and then comment on why I don't believe we need to worry too much about long lists of disjuncts in IBR.

Some correct predictions indeed scale up unhampered. For instance, plain disjunctions with more than two disjuncts receive intuitively correct predictions for any number of disjuncts (if interpreted in epistemic context models, that is). This also implies that the IBR model has no difficulties in explaining the intuitively correct implicatures for sentences like (98) that Chierchia (2004) identified as problematic for more standard Gricean accounts of quantity implicature (see Franke 2009: §3, for treatment of this case in IBR).[31]

(98) Kai had the broccoli or some of the peas.

31 Thanks to an anonymous reviewer for raising this issue.

However, other intuitive results do not scale up easily. Without additional assumptions in the context model, the IBR model predicts that $\Diamond(A \vee B \vee C)$, for instance, is either a surprise message (in the sequence starting with R_0) or interpreted as "exactly two options out of $\{A, B, C\}$ are allowed" (in the sequence after S_0). These predictions are clearly nonsensical. With two plausible extra assumptions, however, the IBR model predicts the correct reading for $\Diamond(A_1 \vee \cdots \vee A_n)$ for any $n > 2$. For that we only need to assume that (i) disjunctions are costly proportional to the number of disjuncts and that (ii) states are more likely the fewer alternatives they make true (e.g., the fewer alternatives are permitted). Similar assumptions would also be needed for scaling-up the derivation of SDA.[32]

Be that as it may, perhaps we should not worry about this too much. As we have just seen, context models and pragmatic reasoning can get *enormously complicated* as n grows bigger.[33] However, it is not necessary — even from a very conservative Gricean point of view — to assume that a rationalistic explanation of a general inference pattern cannot also make recourse to the idea that language users reason in detail about simple cases and *extrapolate* to complex cases. This extrapolation, this generalization from simple to complex, is not an explicit part of the IBR model, but it would make a feasible Gricean companion to the IBR model. Seen from this perspective, we could say that IBR does a fair part of the ground work, but needs to be backed up with an account of pattern recognition and carry-over inferences of a general kind.

Spurious state distinctions. Spurious state distinctions may hamper the derivation of inferences that the IBR model ideally *should* be able to deal with. For example, Chemla (2009) discusses the case in (99) which is problematic for standard Gricean approaches (cf., Geurts & Pouscoulous 2009b, van Rooij 2010).

(99) Everybody is allowed to take an apple or a pear. $\forall x\colon \Diamond(Ax \vee Bx)$

(100) a. Everybody is allowed to take a pear. $\forall x\colon \Diamond Ax$

b. Everybody is allowed to take an apple. $\forall x\colon \Diamond Bx$

32 Another conceivable solution would be to allow asymmetries in the set of alternatives (see also Footnote 15). This, however, is beyond the scope of this paper.

33 Notice that this as such is not a problem for the IBR model alone. It does not get easier for exhaustification-based approaches either to calculate predictions for larger sets of alternatives.

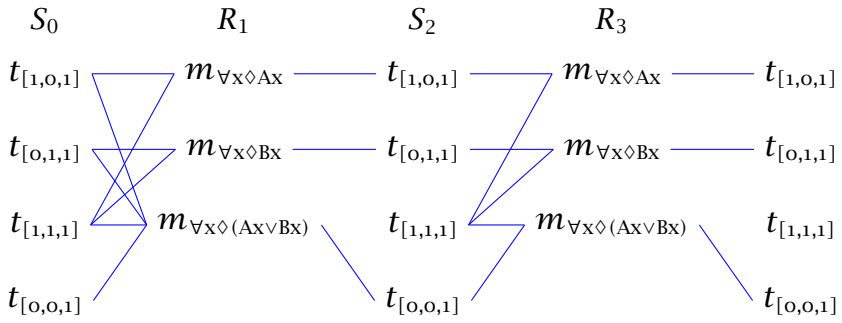


Figure 17 Predictions for example (99) in full context model

It seems fairly intuitive that, at least for deontic modality, the implicatures in (100) should be derivable from (99).[34] But can the IBR model do it?

If we assume that (100) are the alternatives to (99), the base-level interpretation game for this case would distinguish four states:

(101)

	$m_{\forall x \Diamond Ax}$	$m_{\forall x \Diamond Bx}$	$m_{\forall x \Diamond (Ax \vee Bx)}$
$t_{[1,0,1]}$	1	0	1
$t_{[0,1,1]}$	0	1	1
$t_{[1,1,1]}$	1	1	1
$t_{[0,0,1]}$	0	0	1

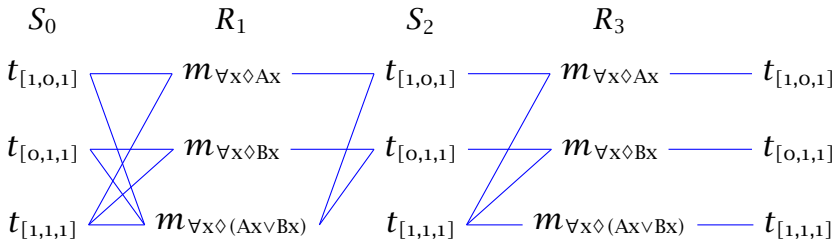
With this the IBR reasoning is as exemplified in Figure 17: we derive that (99) is associated with the state $t_{[0,0,1]}$, which means that we derive the implicature that the group is mixed: some folks may take an apple but no pear, some others may take a pear but no apple. This is not the intuitively correct prediction.

However, in this particular case, a reasonable means of pruning the context model is ready-at-hand. Notice that the state $t_{[0,0,1]}$ *itself* may be fairly unreasonable: it may be deemed very unlikely in context, or even completely ruled out as a live possibility. This is not necessarily a technical trick: it may be a standing presupposition in a natural context of utterance which we construct for sentences like (99) that everybody has *equal rights*. We could implement this assumption in either of two ways: either we could assign to state $t_{[0,0,1]}$ an extremely low, but positive probability, or we could

34 Geurts & Pouscoulous (2009b) argue that implicatures of this kind are not as natural for existential operators other than deontic modals. This observation fits in nicely with the account given here.

scrap it entirely from the context representation. The effect is the same, but the latter affords less computation to spell out the example. Without $t_{[0,0,1]}$ we simply get:

(102)



This is the desired implicature, which we derive — in Gricean fashion — whenever an assumption of "group homogeneity" is feasible simply by pruning the set of state distinctions.

An assumption of "group homogeneity" also deals with the negative counterpart of (99), as discussed by Chemla (2009) as well.[35] The inference in question — constructed here after Chemla's example (15) — is one from (103) to (104) in a context where (105) holds.

(103) Nobody must take an apple and a pear. $\neg \exists x \colon \Box(Ax \wedge Bx)$

(104) a. Everybody may take an apple. $\forall x \colon \Diamond Ax$

b. Everybody may take a pear. $\forall x \colon \Diamond Bx$

(105) Everybody must take an apple or a pear. $\forall x \colon \Box(Ax \vee Bx)$

Indeed, if we assume that (100) are the alternatives to (99), then it's natural to assume that (106) are the relevant alternatives to (103). In that case, the IBR model predicts this inference if (and only if) we assume "group homogeneity" in the same way as before.

(106) a. Nobody must take an apple. $\neg \exists x \colon \Box Ax$

b. Nobody must take a pear. $\neg \exists x \colon \Box Bx$

It is not my intention to defend this analysis come-what-may. I consider it only to demonstrate that it may occasionally be beneficial and reasonable to prune or alter the context model in certain ways. But saying this much

35 I am grateful to an anonymous reviewer for redirecting my attention to this case. The same reviewer also points out that an assumption of "group homogeneity" is problematic because it is liable to overgenerate. I agree with the verdict, not quite the reason. The problem is not overgeneration as such but the lack of an independent standard when this assumption is warranted and when it is not. In any case, this remains an interesting issue for future consideration which I have to leave open here (but see also Chemla 2009: §3.3).

is still far removed from the necessary principled and general specification of exactly which alterations and extra assumptions should be added under exactly which circumstances — an open end for further scrutiny.

10 Reflection & comparison

Before concluding, we should take a brief moment of reflection on this paper's proposal, some of its further implications and its relation to other like-minded models.

Relation to other game theoretic approaches. The approach taken in this paper differs from previous game theoretic approaches to Gricean pragmatics in several respects. The main conceptual difference is that the present approach does not employ equilibrium notions directly (e.g., Parikh 1991, 2001, de Jager & van Rooij 2007, van Rooij 2008), but that it rather selects equilibria indirectly based on an *epistemic solution concept* that is formulated explicitly in terms of the mental states of the players. This approach solves the problem of equilibrium selection that we encountered in Section 7 by implementing semantic meaning as a focal element within pragmatic reasoning. This demonstrably led to better empirical predictions.

Another major difference to other game theoretic approaches is that this paper includes a fully systematic derivation of the game model. The assumptions that motivated the construction were the standard Gricean ones of (i) relevance of information and (ii) cooperativity — nothing more, nothing less. This speaks directly to the often articulated criticism that game theoretic explanations of pragmatic facts are trivially omnipotent, since game parameters only need to be tweaked skillfully enough for any desirable prediction to be obtained. Of course, one must make assumptions about (i) the set-up of the game and (ii) the nature of pragmatic reasoning. This is not a problem as such. Auxiliary assumptions feed into most if not all explanatory theories. The crucial issue really is how *transparent* these assumptions are, and how well they can be made *plausible* either by the general conventions of a scientific community, or, even better, based on domain-independent principles. As for both transparency and domain-independent plausibility, I believe that the IBR model presented here outperforms its competitors easily. Especially, due to its appeal to the logic of folk-psychological explanations every lay person can assess whether the underlying assumptions — such as unbiased beliefs, the implementation of sender competence etc. — are

reasonable or not. I consider this an undeniable methodological advantage *even if* — as I do here — the mental operations in question are not necessarily considered psychologically real.

Relation to exhaustive interpretation. The epistemic approach to rationalizing quantity reasoning also has very interesting theoretical spin-offs. The IBR model shows that in order to compute simple quantity implicatures we do not necessarily need to assume that the speaker adheres to the Maxims of Conversation as we would in traditional Gricean manner. We also do not necessarily need to appeal to the Speaker Quantity Principle in (3) to motivate exhaustive interpretation in terms of minimal models. For clarity: we *can*, but it is *not necessary*.[36] Here is why.

Look at the way R_1 is defined in the light IBR system: for any message m, R_1 selects those states in which m is true and in which fewest alternatives are true (from the set of states where m is true):

(107) $R_1(m) = \{t \in [\![m]\!] \mid \neg\exists t' \in [\![m]\!]\colon |R_0^{-1}(t')| < |R_0^{-1}(t)|\}$.

This is *almost* what exhaustive interpretation in terms of minimal models — as defined in (38) and (41) — does too. To see this, consider the case of base-level interpretation games and EXH_{MM} as defined in (38). The minimal worlds according to EXH_{MM} are obtained by the partial order $<_{\text{ALT}}$ defined in terms of set inclusion of $A(w) = \{A \in \text{ALT} \mid w \in A\}$, whereas the minimal states — sets of worlds — that R_1 chooses are obtained by a total order defined in terms of $|A(t)|$ where $A(t) = \{A \in \text{ALT} \mid t \subseteq \text{ALT}\}$. It is obvious that all worlds in every state in $R_1(m)$ will be in EXH_{MM}:

Fact 1. Base-level exhaustive interpretation in terms of minimal models, is entailed by level-1 receiver interpretation: $\bigcup R_1(m_S) \subseteq \text{EXH}_{\text{MM}}(S)$.

The reverse, however, is not necessarily the case: it may well be that two worlds $w, v \in S$ are both minimal with respect to $<_{\text{ALT}}$, while w makes more alternatives true than v. In that case, w would not occur in any state in $R_1(m_S)$, but it would be in $\text{EXH}_{\text{MM}}(S)$. However under a natural condition that the set of alternatives is in a sense "*homogeneous*", it is actually the

36 For instance, a different rationalization of epistemic exhaustive interpretation in terms of minimal models is given by de Jager & van Rooij (2007). This characterization, however, relies on a different set-up of the signaling game and further strong assumptions about the set ALT that are incompatible with some of the central examples we look at here.

case that exhaustive interpretation coincides with level-1 interpretation.[37] Analogous remarks apply to epistemic interpretation via $\text{EXH}^{\text{GE}}_{\text{MM}}$ defined in (41), and level-1 receiver behavior in epistemic interpretation games without competence assumption.

Whether inclusion or identity, this result is actually quite remarkable. Contrary to the wide-spread conviction that exhaustive interpretation is licensed by something like the Speaker Quantity Principle, it actually *does not require* the assumption of a rational speaker to justify exhaustive interpretation as rational. More strongly even, it is not necessarily the case that EXH_{MM} also coincides with R_2, which captures the interpretation of a rational hearer who believes in a speaker that prefers more informative utterances over less informative ones (see, for example, Figure 7 where $R_1 \neq R_2$). But that means that, given the assumptions of the IBR model, exhaustive interpretation in terms of minimal models is best characterized as the interpretation of a rational hearer who believes that the speaker randomly says something true, not that of a (rational) hearer who believes the speaker conforms to a Gricean Quantity Maxim.

Taking the step from exhaustive interpretation to quantity implicatures in general, the IBR model shows that simple quantity inferences do not need the assumption of a maximally informative speaker, but only require rational interpretation based on an understanding of conventional semantic meaning *plus* a frequency-based inference. For example, a level-1 receiver's interpretation of m_{some} in the game from Figure 4 is $t_{\exists\neg\forall}$, because if we assume truthfulness, a speaker in $t_{\exists\neg\forall}$ is twice as likely to have emitted m_{some} than a speaker in $t_{\forall}$. Technically, this is just the effect of rational belief formation by Bayesian conditionalization (see Appendix B.1). What this means is that simple quantity implicatures can be rationalized by a simple kind of "frequency reasoning" based on the distributional information in a set of meaningful alternatives alone. In other words, the IBR model here gives rise to a rather iconoclastic conclusion: some *quantity* implicatures follow from *quality* alone when we assume that the hearer interprets rationally.

37 We could spell out "homogeneity", for example, as the requirement that there be an isomorphism between any two maximal consistently-excludable sets in $\text{Max-CE}(S, \text{ALT})$ that preserves entailment relations. But it is not essential to the purpose of this article to spell out this condition — or any other sufficient condition on ALT — in more detail. It suffices to note that all examples discussed here indeed make it true that $\text{EXH}_{\text{MM}}(S) = \bigcup R_1(m_S)$. It should also be added that the fact that $<_{\text{ALT}}$ is a partial ordering does some good pragmatic work in the framework of Schulz & van Rooij (2006) when we consider more fine-grained semantic entities as in dynamic semantics, where we also keep track of discourse referents.

11 Conclusion

Taking stock, this paper has offered a general game theoretic model of quantity implicature calculation. The model consists of two parts: (i) a general procedure with which to construct interpretation games as models of the context of utterance from a set of alternative sentences, and (ii) a step-by-step reasoning process that selects the pragmatically feasible play in these games. This approach deals uniformly with a variety of quantity implicatures, and is also versatile enough to make general and yet flexible predictions about different strengths of epistemic quantity implicatures.

The model's predictions are given by fixed points of ever more sophisticated theory-of-mind reasoning. Central results established that fixed points are always reached for interpretation games, and that these are equilibria. The paper also showed that we can compute IBR reasoning by a manageable algorithm, if parameters are adequately chosen. Of course, generally a game theoretic approach to pragmatic reasoning is much more powerful than the specialized model of this paper. The present model would straightforwardly extend also to cases, for instance, where the speaker's and the hearer's interests are in conflict, where message costs are severe, etc.

The model given in this paper is superficially very similar to (iterated) exhaustive interpretation. It subsumes exhaustification in terms of minimal models as a special case in all examples that we looked at in this paper, namely as rational interpretation of a hearer who merely takes the distributional information given by the semantic meanings of a set of alternative expressions into account. The model, however, deviates from the predictions of iterated applications of EXH_{MM} in that it is not monotonic: unlike the latter, the former reconsiders all options (formulations or interpretations) anew at each iteration step. This way the "pragmatically proper" mappings between formulations and interpretations may be established at later iterations, even if excluded at earlier steps.

But whereas the IBR model bears a clear relationship to iterated exhaustification based on minimal models, there is also an uncomfortable gap in the picture sketched in this paper. The IBR model does not relate in the same obvious way to exhaustive interpretation in terms of innocent exclusion. It remains unclear from the present perspective if and how this interpretation operation could be explained as a rational inference or, more generally, as the outcome of some process optimizing speaker's and hearer's interests in successful communication. Presently, however, I must leave this too for future consideration.

A Comparison of exhaustivity operators

This section compares the semantic (minimal-models) approach to exhaustive interpretation with the syntactic (innocent-exclusion) approach (see Section 4 for definitions). Three results characterize the general relationship between these approaches: (i) Fact 2 states that EXH_{MM} is insensitive to certain variations in the set ALT, (ii) Fact 3 establishes that the interpretation selected by EXH_{MM} always entails that selected by EXH_{IE}, and (iii) Fact 4 gives sufficient and necessary conditions on ALT for identity of predictions.

To start with, here is a simple example that shows that the operators EXH_{MM} and EXH_{IE} are not generally equivalent. Consider a disjunction with two (logically independent) disjuncts and two candidate sets of alternatives.

(108) A or B

(109) a. $\text{ALT}_1 = \{A,\ B,\ A \vee B\}$
b. $\text{ALT}_2 = \text{ALT}_1 \cup \{A \wedge B\}$

Let's calculate the predictions for both approaches, starting with the approach in terms of minimal models. We only need to consider three types of worlds that are true in S: w_A where A is true and B is false, w_B where B is true and A is false and w_{AB} where both A and B are true. The minimal worlds in this triple with respect to both $<_{\text{ALT}_1}$ and $<_{\text{ALT}_2}$ are all worlds of type w_A or of type w_B. In other words, the additional conjunctive alternative in ALT_2 does not influence the ordering on possible worlds. Hence, the predictions of the minimal-models approach are the same for both sets of alternatives:

(110) $\text{EXH}_{\text{MM}}(A \vee B, \text{ALT}_1) = \text{EXH}_{\text{MM}}(A \vee B, \text{ALT}_2) = \{w_A, w_B\}$.

This is different for the innocent-exclusion approach. Consider first ALT_1. The maximal consistently excludable sets are $\{A\}$ and $\{B\}$, but their intersection is empty, so that

(111) $\text{EXH}_{\text{IE}}(A \vee B, \text{ALT}_1) = A \vee B = \{w_A, w_B, w_{AB}\}$.

Considering instead ALT_2, the maximal consistently excludable sets are $\{A,\ A \wedge B\}$ and $\{B,\ A \wedge B\}$. These have a non-empty intersection so that

(112) $\text{EXH}_{\text{IE}}(A \vee B, \text{ALT}_2) = \{w_A, w_B\}$.

This is equivalent to the prediction of the semantic approach.

Taken together, the presence or absence of the conjunctive alternative matters to the syntactic, but not to the semantic approach. The latter predicts *as if* the conjunctive alternative was given. Schematically:

(113) $\text{Exh}_{\text{MM}}(\text{Alt}_1) = \text{Exh}_{\text{MM}}(\text{Alt}_2) = \text{Exh}_{\text{IE}}(\text{Alt}_2) \subset \text{Exh}_{\text{IE}}(\text{Alt}_1)$.

We can generalize the point of this example. Indeed, the predictions of the semantic approach are necessarily equivalent under a certain variance in the set Alt; the syntactic approach does not have this invariance property. Notice that in the example above, the orderings $<_{\text{Alt}_1}$ and $<_{\text{Alt}_2}$ were identical because each world that assigns truth values to propositions A and B *thereby* also assigns a truth value to $A \wedge B$. Let's say that a proposition A is *truth-determined* by a set of propositions X if the truth value of A is completely determined by any truth value assignment to all members of X. The ordering $<_X$ is invariant under addition or removal of truth-determined alternatives in the following sense:

Fact 2. If A is truth-determined by X, then $<_X = <_{X \cup \{A\}}$.

This means that Exh_{MM} is less sensitive to the exact specification of the alternatives: if certain propositions A and B are alternatives in Alt, then Exh_{MM} also, as it were, implicitly considers the conjunctive alternative $A \wedge B$. In contrast, Exh_{IE} does not.

This suggests that pragmatic interpretation in terms of Exh_{MM} is always included — but not necessarily strictly — in the pragmatic interpretation in terms of Exh_{IE}. This is borne out in general:

Fact 3. For any S and Alt, we have $\text{Exh}_{\text{MM}}(S, \text{Alt}) \subseteq \text{Exh}_{\text{IE}}(S, \text{Alt})$.

Proof. Take an arbitrary $w \in \text{Exh}_{\text{MM}}(S, \text{Alt})$. This means that w is minimal in S with respect to ordering $<_{\text{Alt}}$, which in turn means that w makes a maximal number of alternatives in Alt false. In other words, there is an $\mathcal{A} \in \text{Max-CE}(S, \text{Alt})$ such that $w \in S \wedge \bigwedge_{A \in \mathcal{A}} \neg A$. But then w is also in any proposition $S \wedge \bigwedge_{A \in \mathcal{B}} \neg A$ for $\mathcal{B} \subseteq \mathcal{A}$. In particular, then, $w \in \text{Exh}_{\text{IE}}(S, \text{Alt})$. □

Next, we would like to know under which circumstances exactly the operators coincide. This, obviously, hinges on the set Alt. While Exh_{MM} rules out all non-minimal worlds according to $<_{\text{Alt}}$, the operator Exh_{IE} does not necessarily rule out all non-minimal worlds, but only those worlds $w \in S$ for

which there is an alternative $A \in \text{ALT}$ such that *all* minimal worlds make A true (resp. false), while w makes A false (resp. true). In other terms, EXH_{IE} is more conservative than EXH_{MM} in that it selects as pragmatic interpretation of S all those worlds in S which cannot be distinguished from the minimal worlds by some alternative in ALT. To formalize this, we introduce a suitable notion of distinguishability: we say that world w is ALT-distinguishable from the set of worlds $X \subseteq W$ iff there is some $A \in \text{ALT}$ such that either all worlds in X make A true while w makes A false, or all worlds in X make A false while w makes A true. If w is ALT-indistinguishable from set X, write $w \sim_{\text{ALT}} X$. These considerations give rise to the following:

Lemma 1. $\text{EXH}_{\text{IE}}(S, \text{ALT}) = \{w \in S \mid w \sim_{\text{ALT}} \text{EXH}_{\text{MM}}(S, \text{ALT})\}$

In other terms, we can characterize EXH_{IE} semantically as the closure of EXH_{MM} under ALT-indistinguishability.

So when exactly does a set ALT have the property that the minimal worlds according to $<_{\text{ALT}}$ are closed under ALT-indistinguishability? To state sufficient and necessary conditions for this formally, define for all $w \in \text{EXH}_{\text{MM}}$ that $\text{A}(w)$ is the unique strongest proposition in $\text{ALT} \cup \{S\}$ that is true in w. With this, we can state the following sufficient and necessary condition on ALT for equivalence of EXH_{MM} and EXH_{IE}.

Fact 4. $\text{EXH}_{\text{MM}}(S, \text{ALT}) = \text{EXH}_{\text{IE}}(S, \text{ALT})$ iff for all $w, w' \in \text{EXH}_{\text{MM}}(S, \text{ALT})$ there is an alternative $A \in \text{ALT}$ such that the conjunction of $\text{A}(w)$ and $\text{A}(w')$ entails A.

Proof. With Lemma 1 it suffices to show that the right hand side of Fact 4 is equivalent to:

(114) $\text{EXH}_{\text{MM}}(S, \text{ALT}) = \{w \in S \mid w \sim_{\text{ALT}} \text{EXH}_{\text{MM}}(S, \text{ALT})\}$.

This is so because, firstly, if for all worlds w, w' in $\text{EXH}_{\text{MM}}(S, \text{ALT})$ there is an A with the designated property, then we *can* distinguish all non-minimal worlds from the minimal ones. Secondly, if for a given pair of worlds w, w' in $\text{EXH}_{\text{MM}}(S, \text{ALT})$ there is no such A with the designated property, then there is a world w^* which makes both $A(w)$ and $A(w')$ true and all other $A \in \text{ALT}$ false. This w^* is not minimal, but it is also not ALT-distinguishable from the set $\{w, w'\}$ and therefore also not from $\text{EXH}_{\text{MM}}(S, \text{ALT})$. □

B IBR reasoning — formal background

The goal of this section is to fill in some of the formal details of IBR reasoning that the main paper skipped in the interest of readability. Section B.1 spells out the players' beliefs and rational behavior in more traditional terms and Section B.2 shows how this derives the "light system" of Section 8. Section B.3 discusses an example in the "heavy system" of Section B.1 for clarity, and Section B.4 finally states and proves some general properties of the IBR model.

B.1 The heavy system

As before, player types are defined via sets of pure strategies. The definition of level-0 players is exactly as in the light system. The crucial detail that the main paper blended out is in the definition of rational behavior given a certain belief. To fill in this detail, let's focus on the sender side first. If, for instance, a level-$(k+1)$ sender has an unbiased belief in a given set R_k of pure strategies, then we can derive from that the sender's so-called *behavioral beliefs* — formally a function in $(\Delta(A))^M$ — that represent how likely the sender believes it is that a given action is played in response to a given message. For unbiased beliefs in R_k this is not difficult to compute: the probabilistic sender belief that m is answered by a is simply the proportion of receiver strategies in R_k that map m to a. Abusing notation, we write:

$$R_k(m,a) = \frac{|\{r \in A^M \mid r(m) = a\}|}{|R_k|}\,. \tag{115}$$

Notice that since the only thing that the sender does not know when it comes to her making a move is the receiver's behavior, the sender's game relevant uncertainty is entirely captured by a behavioral receiver strategy. Using the standard definition of rationality as maximization of expected utility, we define the set of all *best responses* to the unbiased belief R_k as:

$$\mathrm{BR}(R_k) = \left\{ s \in M^T \mid \forall t\colon s(t) \in \arg\max_{m \in M} \sum_{a \in A} R_k(m,a) \times \mathrm{U}_S(t,m,a) \right\}. \tag{116}$$

Adopting the TCP assumption, we then define the behavior of a level-$(k+1)$ sender as a best response in $\mathrm{BR}(R_k)$ that, if possible, respects semantic meaning:

$$S_{k+1} = \{s \in \mathrm{BR}(R_k) \mid \forall t\ (\exists s' \in \mathrm{BR}(R_k)\ t \in [\![s'(t)]\!]) \rightarrow t \in [\![s(t)]\!]\}\,. \tag{117}$$

A largely parallel definition yields higher level receiver types. However, since the receiver is uncertain about, not only what the sender does, but also about what state the sender has observed, the definition of the receiver's beliefs are slightly more complicated. As before, though, we can define the receiver's *behavioral beliefs* — formally: a function in $(\Delta(M))^T$ — given an unbiased belief in S_k (with harmless overload of notation) as:

$$S_k(t,m) = \frac{|\{s \in M^T \mid s(t) = m\}|}{|S_k|} . \tag{118}$$

These beliefs, however, only indirectly feed into the definition of receiver rational behavior, because it is primarily the receiver's *posterior beliefs* that decide what counts as a rational choice. The posterior beliefs $\mu \in (\Delta(T))^M$ specify how likely the receiver considers a state after observing a given message. Obviously, posterior beliefs should be a function of prior and behavioral beliefs wherever possible. The normatively correct way of forming posterior beliefs is by *Bayesian conditionalization*. We say that μ is *consistent* with Pr and S_k iff for all t and m for which there is a state t' such that $S_k(m|t') \neq 0$ we have:

$$\mu(t|m) = \frac{\Pr(t) \times S_k(m|t)}{\sum_{t' \in T} \Pr(t') \times S_k(m|t')} . \tag{119}$$

Consistency effectively demands conservative belief dynamics: wherever possible Bayesian conditionalization computes backward the *likelihood* for each state t that an observed message m was sent in t given t's prior probability and the probability with which m was expected to be sent in t.

With this we can define what a best response to a posterior belief μ is:

$$\mathrm{BR}(\mu) = \left\{ r \in A^M \mid \forall m\colon\ r(m) \in \arg\max_{a \in A} \sum_{t \in T} \mu(t|m) \times \mathrm{U}_R(t,m,a) \right\} . \tag{120}$$

This gives us the proper definition of a best response to an unbiased belief in S_k:

$$\mathrm{BR}(S_k) = \{\mathrm{BR}(\mu) \mid \mu \text{ is consistent with Pr and } S_k\} . \tag{121}$$

Under the TCP assumption, level-$(k+1)$ receiver behavior is then defined as:

$$R_{k+1} = \{r \in \mathrm{BR}(S_k) \mid \forall m\ (\exists r' \in \mathrm{BR}(S_k)\ r'(m) \in [\![m]\!]) \rightarrow r(m) \in [\![m]\!]\} . \tag{122}$$

B.2 Journey from heavy to light

Reasoning with flat priors. To see how these latter definitions give rise to the light system of Section 8, first recapitulate the inductive step of the definition of the light system with flat priors:

$$\check{S}_{k+1}(t) = \begin{cases} \arg\min_{m \in \check{R}_k^{-1}(t)} |\check{R}_k(m)| & \text{if } \check{R}_k^{-1}(t) \neq \emptyset \\ \check{S}_0(t) & \text{otherwise.} \end{cases} \tag{123}$$

$$\check{R}_{k+1}(m) = \begin{cases} \arg\min_{t \in \check{S}_k^{-1}(m)} |\check{S}_k(t)| & \text{if } \check{S}_k^{-1}(m) \neq \emptyset \\ \check{R}_0(m) & \text{otherwise.} \end{cases} \tag{124}$$

We can then state this section's first main result as follows:

Theorem 1. The light system is an equivalent reformulation of the previous heavy system, i.e., $R_k = \check{R}_k$ and $S_k = \check{S}_k$ for all k, if we assume that the game model satisfies conditions:

C1: $T = A$;

C2: $\mathrm{U}_{S,R}(t, m, a) = 1$ if $t = a$ and 0 if $t \neq a$;

C3: $\Pr(t) = \Pr(t')$ for all t, t'.

Towards a proof, first formulate and prove the following:

Lemma 2. Under the conditions C1 and C2 truth is preserved by all types: $S_k(t) \subseteq [\![t]\!]^{-1}$ and $R_k(m) \subseteq [\![m]\!]$ for all k.

Proof of Lemma 2. By induction. The base case is trivial. So suppose that $S_k(t) \subseteq [\![t]\!]^{-1}$ for all t, and show $R_{k+1}(m) \subseteq [\![m]\!]$, for all m. First, take the case $S_k^{-1}(m) \neq \emptyset$, i.e., a non-surprise message m. By inductive hypothesis $S_k^{-1}(m) \subseteq [\![m]\!]$, and so together with conditions C1 and C2:

$$R_{k+1}(m) = \arg\max_{t \in T} \mu_{k+1}(t|m) \subseteq S_k^{-1}(m) \subseteq [\![m]\!]. \tag{125}$$

In case of a surprise message with $S_k^{-1}(m) = \emptyset$, any state is a best response, and so by TCP assumption $R_{k+1}(m) \subseteq [\![m]\!]$.

The induction step for the sender is almost identical. Suppose that $R_k(m) \subseteq [\![m]\!]$ for all m, and show $S_{k+1}(t) \subseteq [\![t]\!]^{-1}$ for all t. First, take the case $R_k^{-1}(t) \neq \emptyset$. By induction hypothesis, C1 and C2:

$$S_{k+1}(t) = \arg\max_{m \in M} \sum_{a \in A} R_k(m, a) \times \mathrm{U}_S(t, m, a) \subseteq R_k^{-1}(t) \subseteq [\![t]\!]^{-1}. \tag{126}$$

In case $R_k^{-1}(t) = \emptyset$, any message maximizes expected utility given R_k, so that by TCP assumption $S_{k+1}(t) \subseteq [\![t]\!]^{-1}$. □

Proof of Theorem 1. By induction. As the base case is trivial, assume first that $R_k = \breve{R}_k$ and show that $S_{k+1} = \breve{S}_{k+1}$. By definition:

$$S_{k+1} = \{s \in \mathrm{BR}(R_k) \mid \forall t \ (\exists s' \in \mathrm{BR}(R_k) \ t \in [\![s'(t)]\!]) \rightarrow t \in [\![s(t)]\!]\} \ . \tag{127}$$

First, take a state for which $R_k^{-1}(t) \neq \emptyset$. From Lemma 2, we know that $R_k(m) \subseteq [\![m]\!]$, so that:

$$S_{k+1}(t) = \arg\max_{m \in M} \sum_{a \in A} R_k(m,a) \times \mathrm{U}_S(t,m,a) \ . \tag{128}$$

This expected utility boils down to the following under assumptions C1 and C2:

$$\begin{aligned} \sum_{a \in A} R_k(m,a) \times \mathrm{U}_S(t,m,a) &= \sum_{t' \in T} R_k(m,t') \times \mathrm{U}_S(t,m,t') \\ &= \begin{cases} \frac{1}{|R_k(m)|} & \text{if } t \in R_k(m) \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{129}$$

From the last two equations, $S_{k+1}(t) = \breve{S}_{k+1}(t)$ follows. If, on the other hand, $R_k^{-1}(t) = \emptyset$, then any message will maximize expected utility, and the speaker will, by TCP assumption, send arbitrarily any true message, so that $S_{k+1}(t) = [\![t]\!]^{-1} = \breve{S}_{k+1}(t)$.

Finally, assume that $S_k = \breve{S}_k$ and show that $R_{k+1} = \breve{R}_{k+1}$. For surprise messages with $S_k(m)^{-1} = \emptyset$, any state maximizes expected utility given some consistent belief in S_k. In that case, by TCP assumption $R_{k+1}(m) = \breve{R}_{k+1}(m)$. So, suppose m is not a surprise, i.e., $S_k(m)^{-1} \neq \emptyset$. Then Bayesian conditionalization gives a unique $\mu_{k+1}(\cdot|m)$ for which by C1 and C2:

$$R_k(m) = \arg\max_{t \in T} \mu_{k+1}(t|m) \ . \tag{130}$$

To solve this maximization, calculate:

$$\begin{aligned} & & & \mu_{k+1}(t_1|m) > \mu_{k+1}(t_2|m) \\ & & \text{iff} \quad & \frac{\Pr(t_1) \times S_k(t_1,m)}{\sum_{t' \in T} \Pr(t') \times S_k(t',m)} > \frac{\Pr(t_2) \times S_k(t_2,m)}{\sum_{t' \in T} \Pr(t') \times S_k(t',m)} \\ & & \text{iff} \quad & \Pr(t_1) \times S_k(t_1,m) > \Pr(t_2) \times S_k(t_2,m) \\ & \text{(from C3)} & \text{iff} \quad & S_k(t_1,m) > S_k(t_2,m) \\ & & \text{iff} \quad & (m \in S_k(t_1) \text{ and } m \notin S_k(t_2)) \text{ or} \\ & & & (m \in S_k(t_1) \text{ and } m \in S_k(t_2) \text{ and } |S_k(t_1)| < |S_k(t_2)|) \ . \end{aligned} \tag{131}$$

With this, it is clear that $R_k(m) = \breve{R}_k(m)$. □

Reasoning with near-flat priors. If priors are not flat but differences are small enough, the receiver's reasoning can be describe as follows:

Theorem 2. Let $t_{\max}$ and $t_{\min}$ be the most and least likely states of a signaling game that satisfies conditions C1 and C2 of Theorem 1, plus

$$\frac{\Pr(t_{\min})}{\Pr(t_{\max})} < \frac{|M|-1}{|M|}\,. \tag{132}$$

Then for all $k \geq 0$:

$$R_{k+1}(m) = \left\{ t \in \breve{R}_{k+1}(m) \mid \neg\exists t' \in \breve{R}_{k+1}(m)\colon\ \Pr(t') > \Pr(t) \right\}\,. \tag{133}$$

Proof. We need to show that the given condition implies that the only case where prior probabilities ever make a difference between the posterior likelihood of two states t and t' given some message m is when $S_k(t,m) = S_k(t',m)$. To show this it suffices to look at $t_{\min}$ and $t_{\max}$ and the "worst case" where $S_k(t_{\min},m) > S_k(t_{\max},m)$. We need to show that

$$\Pr(t_{\min}) \times S_k(t_{\min},m) > \Pr(t_{\max}) \times S_k(t_{\max},m) \tag{134}$$

even if the difference between $S_k(t_{\min},m)$ and $S_k(t_{\max},m)$ is as small as it can possibly get. Since we are dealing with interpretation games, this "worst case" is when $S_k(t_{\min},m) = \frac{1}{|M|-1}$ and $S_k(t_{\max},m) = \frac{1}{|M|}$. But if the priors satisfy the required condition, then the above holds true. □

B.3 The basic free choice implicature in the heavy system

To better understand the probabilistic reasoning of the "heavy system", in particular (i) the effect of Bayesian conditionalization on pragmatic interpretation and (ii) the notions "surprise message" and "uninducible interpretation", let us compute the predictions for the game in Figure 5. The unsophisticated receiver behavior in this game is given by the semantic meaning of messages only:

$$R_0 = \left\{ \begin{array}{lcl} m_{\Diamond \mathrm{A}} & \mapsto & t_{\mathrm{A}}, t_{\mathrm{AB}} \\ m_{\Diamond \mathrm{B}} & \mapsto & t_{\mathrm{B}}, t_{\mathrm{AB}} \\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} & \mapsto & t_{\mathrm{A}}, t_{\mathrm{B}}, t_{\mathrm{AB}} \end{array} \right\}\,. \tag{135}$$

This defines the level-1 sender's unbiased behavioral belief: for example, S_1 believes that if she sends $m_{\Diamond \mathrm{A}}$ the receiver will not choose interpretation t_{B} at all, but may choose t_{A} or t_{AB} with equal probability. What is rational behavior under this belief? This can be calculated along the above definitions but it is also intuitively appreciated that, for example, in state t_{A} the only rational choice given this belief is to send $m_{\Diamond \mathrm{A}}$: $m_{\Diamond \mathrm{A}}$ has a probability $1/2$ chance of inducing the correct response, while $m_{\Diamond(\mathrm{A} \vee \mathrm{B})}$ has a probability $1/3$ chance, and $m_{\Diamond \mathrm{B}}$ will simply never elicit the proper response in the hearer. Similar reasoning establishes:

$$(136) \quad S_1 = \left\{ \begin{array}{lcl} t_{\mathrm{A}} & \mapsto & m_{\Diamond \mathrm{A}} \\ t_{\mathrm{B}} & \mapsto & m_{\Diamond \mathrm{B}} \\ t_{\mathrm{AB}} & \mapsto & m_{\Diamond \mathrm{A}}, m_{\Diamond \mathrm{B}} \end{array} \right\}.$$

It is noteworthy here that $m_{\Diamond \mathrm{A}}$ and $m_{\Diamond \mathrm{B}}$ are the best sender choices in t_{AB}, because under R_0's interpretation each of these messages yields a chance of $1/2$ of successful communication, as opposed to a chance of $1/3$ when sending $m_{\Diamond(\mathrm{A} \vee \mathrm{B})}$. That means that the target message $m_{\Diamond(\mathrm{A} \vee \mathrm{B})}$ will actually be a *surprise message* to R_2, as an unbiased belief in S_1 entails a belief that message $m_{\Diamond(\mathrm{A} \vee \mathrm{B})}$ never gets sent. It is crucial to note here that Bayesian conditionalization — as it is needed to compute consistent posterior beliefs, and from there the proper responses of R_2 — does not apply to surprise messages. There is indeed a lot of literature in rational choice theory dealing with how beliefs after surprise messages could or should be formed (cf., Stalnaker 1998). The present IBR model simply predicts that surprise messages could be answered by just any interpretation, were it not for the TCP assumption that, all else equal, players stick to the semantic meaning of messages. So, by TCP assumption, surprise messages are interpreted *literally* like an unsophisticated receiver would. This gives us:

$$(137) \quad R_2 = \left\{ \begin{array}{lcl} m_{\Diamond \mathrm{A}} & \mapsto & t_{\mathrm{A}} \\ m_{\Diamond \mathrm{B}} & \mapsto & t_{\mathrm{B}} \\ m_{\Diamond(\mathrm{A} \vee \mathrm{B})} & \mapsto & t_{\mathrm{A}}, t_{\mathrm{B}}, t_{\mathrm{AB}} \end{array} \right\}.$$

From this point on, the reasoning chain unfolds smoothly. Based on a belief in R_2, the sender will send message $m_{\Diamond(\mathrm{A} \vee \mathrm{B})}$ exactly in state t_{AB}, because in this state this is the only message that has a positive probability

of inducing the right interpretation:

$$(138)\quad S_3 = \left\{ \begin{array}{lcl} t_{\mathrm{A}} & \mapsto & m_{\Diamond \mathrm{A}} \\ t_{\mathrm{B}} & \mapsto & m_{\Diamond \mathrm{B}} \\ t_{\mathrm{AB}} & \mapsto & m_{\Diamond (\mathrm{A} \vee \mathrm{B})} \end{array} \right\}.$$

The only best response to this is for the receiver to interpret as follows:

$$(139)\quad R_4 = \left\{ \begin{array}{lcl} m_{\Diamond \mathrm{A}} & \mapsto & t_{\mathrm{A}} \\ m_{\Diamond \mathrm{B}} & \mapsto & t_{\mathrm{B}} \\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} & \mapsto & t_{\mathrm{AB}} \end{array} \right\}.$$

This sequence of reasoning steps has thereby reached a fixed point.

Next, we should also check the model's predictions starting with an unsophisticated sender. It turns out that this reasoning chain indeed terminates in the exact same fixed point for mostly the same reasons. An unsophisticated sender is given by:

$$(140)\quad S_0 = \left\{ \begin{array}{lcl} t_{\mathrm{A}} & \mapsto & m_{\Diamond \mathrm{A}},\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} \\ t_{\mathrm{B}} & \mapsto & m_{\Diamond \mathrm{B}},\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} \\ t_{\mathrm{AB}} & \mapsto & m_{\Diamond \mathrm{A}},\ m_{\Diamond \mathrm{B}},\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} \end{array} \right\}.$$

In order to compute the set R_1 from this, we need to compute consistent posterior beliefs and then rational responses to these. Formally this is a mild load of work, but the rationale behind this reasoning can also be framed intuitively. For example, a level-1 receiver expects the message $m_{\Diamond \mathrm{A}}$ to be sent with probability $1/2$ in state t_{A}, with probability $1/3$ in state t_{AB} and not at all in state t_{B}. Consequently, by Bayesian conditionalization the state that the receiver thinks is *most likely* after hearing message $m_{\Diamond \mathrm{A}}$ is t_{A}. Since in an interpretation game with its particular payoff structure the rational choice is to go for the most likely interpretation, the receiver will therefore interpret $m_{\Diamond \mathrm{A}}$ as t_{A}. Similar reasoning leads us to verify that:

$$(141)\quad R_1 = \left\{ \begin{array}{lcl} m_{\Diamond \mathrm{A}} & \mapsto & t_{\mathrm{A}} \\ m_{\Diamond \mathrm{B}} & \mapsto & t_{\mathrm{B}} \\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} & \mapsto & t_{\mathrm{A}}, t_{\mathrm{B}} \end{array} \right\}.$$

Two observations are in order here. Firstly, the reasoning that establishes R_1 and S_1 looks very similar, despite the informational asymmetry between

sender and receiver. Indeed, as Theorem 1 established, for interpretation games with flat priors they are *identical*. Secondly, there is an analogue to "surprise messages" also on the sender side. If the sender has an unbiased belief in R_1 then she will believe that it is *impossible* to induce the interpretation t_{AB}: there simply is no message which, according to the sender's beliefs, would have the receiver select this interpretation. By the same reasoning as above the sender would then be indifferent between sending *any* message whatsoever, were it not, again, that we assume with TCP that the sender then *ceteris paribus* prefers to at least send a true message. This derives the player type:

$$S_2 = \left\{ \begin{array}{lcl} t_{\text{A}} & \mapsto & m_{\Diamond \text{A}} \\ t_{\text{B}} & \mapsto & m_{\Diamond \text{B}} \\ t_{\text{AB}} & \mapsto & m_{\Diamond \text{A}},\ m_{\Diamond \text{B}},\ m_{\Diamond (\text{A} \vee \text{B})} \end{array} \right\}. \tag{142}$$

The remaining steps of this IBR reasoning chain are straightforward. We reach the desired fixed point with the following two steps:

$$R_3 = \left\{ \begin{array}{lcl} m_{\Diamond \text{A}} & \mapsto & t_{\text{A}} \\ m_{\Diamond \text{B}} & \mapsto & t_{\text{B}} \\ m_{\Diamond (\text{A} \vee \text{B})} & \mapsto & t_{\text{AB}} \end{array} \right\} \qquad S_4 = \left\{ \begin{array}{lcl} t_{\text{A}} & \mapsto & m_{\Diamond \text{A}} \\ t_{\text{B}} & \mapsto & m_{\Diamond \text{B}} \\ t_{\text{AB}} & \mapsto & m_{\Diamond (\text{A} \vee \text{B})} \end{array} \right\}. \tag{143}$$

It transpires here that the TCP assumption is necessary for assuring truthful production when interpretations are uninducible, as well as literal interpretation of surprise messages. In fact, this is its *only* impact on behavior of IBR types with $k > 0$, and it establishes that *everywhere* in IBR reasoning about interpretation games, players adhere to the conventional meaning of signals (see Lemma 2 and its proof above).

B.4 General results

Here are two more useful general results on IBR reasoning. Firstly, IBR reasoning always reaches a fixed point, if we are dealing with finite games where sender and receiver have aligned preferences.

Theorem 3. For a signaling game of pure cooperation where $\text{U}_S = \text{U}_R$ and where M, A and T are finite, each IBR sequence reaches a fixed point.

Secondly, in interpretation games, this fixed point will always be a *perfect Bayesian equilibrium*, a mild refinement of the above notion of Nash equilibrium (to be defined below).

Theorem 4. Any $\langle S^*, R^* \rangle$ that is a fixed point of an IBR sequence for a signaling game that satisfies conditions C1 and C2 from Theorem 1 gives rise to a perfect Bayesian equilibrium.

The remainder of this section gives the necessary definitions and arguments.

Proof of Theorem 3. Let us define a *signaling game of pure cooperation* as one where sender and receiver utilities are aligned: $\mathrm{U}_S(t,m,a) = \mathrm{U}_R(t,m,a)$ for all t, m, a. For these games, we define the *expected gain* of a pair of strategies σ, ρ as:

$$\mathrm{EG}(\sigma,\rho) = \sum_t \Pr(t) \times \sum_m \sigma(t,m) \times \sum_a \rho(m,a) \times \mathrm{U}(t,m,a)\,. \tag{144}$$

Lemma 3. In a signaling game of pure cooperation the expected gain is monotonically increasing along the IBR sequence, in the sense that for all $i \geq 0$:

(i) $\mathrm{EG}(S_i, R_{i+1}) \leq \mathrm{EG}(S_{i+2}, R_{i+1})$, and

(ii) $\mathrm{EG}(S_{i+1}, R_i) \leq \mathrm{EG}(S_{i+1}, R_{i+2})$.

Proof of Lemma 3. Ad (i). It holds for all t that:

$$S_{i+2}(t) \subseteq \arg\max_{m \in M} \sum_a R_{i+1}(m,a) \times \mathrm{U}_S(t,m,a)\,. \tag{145}$$

This implies that for all $t \in T$:

$$\begin{aligned} &\sum_m S_i(t,m) \times R_{i+1}(m,a) \times \mathrm{U}(t,m,a) \\ &\leq \sum_m S_{i+2}(t,m) \times R_{i+1}(m,a) \times \mathrm{U}(t,m,a)\,. \end{aligned} \tag{146}$$

This, in turn, implies:

$$\begin{aligned} &\sum_t \Pr(t) \sum_m S_i(t,m) \times R_{i+1}(m,a) \times \mathrm{U}(t,m,a) \\ &\leq \sum_t \Pr(t) \sum_m S_{i+2}(t,m) \times R_{i+1}(m,a) \times \mathrm{U}(t,m,a)\,. \end{aligned} \tag{147}$$

And this is equivalent to $\mathrm{EG}(S_i, R_{i+1}) \leq \mathrm{EG}(S_{i+2}, R_{i+1})$.

Ad (ii). Begin by rewriting the statement to be shown:

$$
\begin{aligned}
(148) \qquad & \mathrm{EG}(S_{i+1}, R_i) \leq \mathrm{EG}(S_{i+1}, R_{i+2}) \\
\text{iff} \quad & \sum_t \Pr(t) \times \sum_m S_{i+1}(t,m) \times \sum_a R_i(m,a) \times \mathrm{U}_{S,R}(t,m,a) \leq \\
& \sum_t \Pr(t) \times \sum_m S_{i+1}(t,m) \times \sum_a R_{i+2}(m,a) \times \mathrm{U}_{S,R}(t,m,a) \\
\text{iff} \quad & \sum_t \sum_m \sum_a \Pr(t) \times S_{i+1}(t,m) \times R_i(m,a) \times \mathrm{U}_{S,R}(t,m,a) \leq \\
& \sum_t \sum_m \sum_a \Pr(t) \times S_{i+1}(t,m) \times R_{i+2}(m,a) \times \mathrm{U}_{S,R}(t,m,a) \,.
\end{aligned}
$$

Observe that for messages that surprise R_{i+2} we have $S_{i+1}(t,m) = 0$ for all t, so that the receiver's reception of these messages does not figure in the inequality. Let $M^* = \{m \in M \mid S_{i+1}^{-1}(m) \neq \emptyset\}$ be the set of non-surprise messages under S_{i+1}. The previous statement therefore is equivalent to:

$$
\begin{aligned}
(149) \quad & \sum_t \sum_{m \in M^*} \sum_a \Pr(t) \times S_{i+1}(t,m) \times R_i(m,a) \times \mathrm{U}_{S,R}(t,m,a) \leq \\
& \sum_t \sum_{m \in M^*} \sum_a \Pr(t) \times S_{i+1}(t,m) \times R_{i+2}(m,a) \times \mathrm{U}_{S,R}(t,m,a) \,.
\end{aligned}
$$

Dividing each summand on both sides with a constant $0 \neq c(m) = \sum_{t'} \Pr(t') \times S_{i+1}(t',m)$ for each $m \in M^*$ yields:

$$
\begin{aligned}
(150) \qquad & \sum_t \sum_{m \in M^*} \sum_a \frac{\Pr(t) \times S_{i+1}(t,m)}{c(m)} \times R_i(m,a) \times \mathrm{U}_{S,R}(t,m,a) \leq \\
& \sum_t \sum_{m \in M^*} \sum_a \frac{\Pr(t) \times S_{i+1}(t,m)}{c(m)} \times R_{i+2}(m,a) \times \mathrm{U}_{S,R}(t,m,a) \\
\text{iff} \quad & \sum_t \sum_{m \in M^*} \sum_a \mu_{i+1}(t|m) \times R_i(m,a) \times \mathrm{U}_{S,R}(t,m,a) \leq \\
& \sum_t \sum_{m \in M^*} \sum_a \mu_{i+1}(t|m) \times R_{i+2}(m,a) \times \mathrm{U}_{S,R}(t,m,a) \,.
\end{aligned}
$$

We now see that this inequality holds, because for any $m \in M^*$ we have $R_{i+2}(m) \subseteq \arg\max_a \mu_{i+2}(t|m) \times \mathrm{U}_S(t,m,a)$. □

Proof of Theorem 3. By Lemma 3 we know that expected gain is monotonically increasing. Clearly, $\mathrm{EG}(\cdot)$ is upper-bounded for finite games. Since there are also only finitely many sets of pure strategies (that could constitute types in an IBR sequence), and since the IBR sequence is entirely deterministic, each

sequence must reach a highest value for $\mathrm{EG}(\cdot)$. This entails a fixed point, because from $\mathrm{EG}(S_i, R_{i+1}) = \mathrm{EG}(S_{i+2}, R_{i+1})$, it follows that

$$S_{i+2}(t) \subseteq \arg\max_{m \in M} \sum_a R_{i+1}(m,a) \times \mathrm{U}_S(t,m,a)\,, \tag{151}$$

and this implies that $S_i(t) \subseteq S_{i+2}(t)$ for all t. But for finite set M, this cannot be an infinite sequence with a strict subset relation. □

Proof of Theorem 4. We say that a triple $\langle \sigma, \rho, \mu \rangle \in (\Delta(M))^T \times (\Delta(A))^M \times (\Delta(T))^M$ is a *perfect Bayesian equilibrium* (PBE) iff three conditions hold:[38]

(i) σ is rational given the belief ρ;

(ii) ρ is rational given the belief μ;

(iii) μ is consistent with Pr and the belief σ.

We say that a strategy profile $\langle \sigma, \rho \rangle$ *gives rise to* a PBE iff there is a posterior μ such that $\langle \sigma, \rho, \mu \rangle$ is a PBE.

Lemma 4. If $\langle S^*, R^* \rangle$ is a fixed point of an IBR sequence such that there are no surprise messages under S^*, then $\langle S^*, R^* \rangle$ gives rise to a perfect Bayesian equilibrium.

Proof of Lemma 4. If $\langle S^*, R^* \rangle$ is the fixed point of an IBR sequence, S^* is a best response to the belief R^*. Moreover, if there are no surprise messages under S^*, then there is only one posterior belief μ^* consistent with the given prior and the belief S^*. By definition of IBR types, R^* is a best response to μ^*. Hence, all conditions for perfect Bayesian equilibrium are fulfilled by the triple $\langle S^*, R^*, \mu^* \rangle$. □

Proof of Theorem 4. Given Lemma 4 and its proof, we only need to show that for any fixed point $\langle S^*, R^* \rangle$ there is a posterior μ under which $R^*(m)$ is rational for surprise messages m. Given conditions C1 and C2 this is fulfilled by the unique μ^* which is consistent with S^*, and for which for all surprise messages m we have: $\mu^*(t|m) = |[\![m]\!]|^{-1}$. □

38 Strictly speaking, the notion of rationality of a *probabilistic* strategy under an *arbitrary* behavioral belief has not been defined in this paper, but it bears no surprises. The interested reader is referred to standard textbooks.

References

Allott, Nicholas. 2006. Game theory and communication. In Anton Benz, Gerhard Jäger & Robert van Rooij (eds.), *Game theory and pragmatics*, 123–151. Basingstoke and New York: Palgrave Macmillan.

Alonso-Ovalle, Luis. 2005. Distributing the disjuncts over the modal space. In Leah Bateman & Cherlon Ussery (eds.), *Proceedings of the North East Linguistics Society (NELS) 35*, Amherst, MA: GLSA.

Alonso-Ovalle, Luis. 2008. Innocent exclusion in an alternative-semantics. *Natural Language Semantics* 16(2). 115–128. doi:10.1007/s11050-008-9027-1.

Asher, Nicholas & Daniel Bonevac. 2005. Free choice permission is strong permission. *Synthese* 145(3). 303–323. doi:10.1007/s11229-005-6196-z.

Atlas, Jay David & Stephen Levinson. 1981. It-clefts, informativeness, and logical form. In Peter Cole (ed.), *Radical pragmatics*, 1–61. New York: Academic Press.

Bach, Kent. 2006. The top 10 misconceptions about implicature. In Betty Birner & Gregory Ward (eds.), *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn* (Studies in Language Companion Series 80), 21–30. Amsterdam and Philadelphia, PA: John Benjamins.

Barker, Chris. 2010. Free choice permission as resource-sensitive reasoning. *Semantics & Pragmatics* 3(10). 1–38. doi:10.3765/sp.3.10.

Benz, Anton. 2006. Utility and relevance of answers. In Anton Benz, Gerhard Jäger & Robert van Rooij (eds.), *Game theory and pragmatics*, 195–219. Basingstoke and New York: Palgrave Macmillan.

Benz, Anton, Gerhard Jäger & Robert van Rooij (eds.). 2006. *Game theory and pragmatics*. Basingstoke and New York: Palgrave Macmillan.

Benz, Anton & Robert van Rooij. 2007. Optimal assertions and what they implicate. *Topoi* 26(1). 63–78. doi:10.1007/s11245-006-9007-3.

Camerer, Colin F. 2003. *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Camerer, Colin F., Teck-Hua Ho & Juin-Kuan Chong. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics* 119(3). 861–898. doi:10.1162/0033553041502225.

Chapman, Siobhan. 2005. *Paul Grice, philosopher and linguist*. New York: Palgrave Macmillan.

Chemla, Emmanuel. 2009. Universal implicatures and free choice effects:

Experimental data. *Semantics & Pragmatics* 2(2). 1–33. doi:10.3765/sp.2.2.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In Adriana Belletti (ed.), *Structures and beyond* (The Cartography of Syntactic Structures 3), 39–103. Oxford: Oxford University Press.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2008. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. Unpublished manuscript.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2009. Hurford's constraint and the theory of scalar implicatures. In Paul Egré & Giorgio Magri (eds.), *Presuppositions and implicatures. Proceedings of the MIT-Paris workshop* (MIT Working Papers in Linguistics 60), http://semanticsarchive.net/Archive/mE2OGIzY/HCChierchiaFoxSpector.pdf.

Cho, In-Koo & David M. Kreps. 1987. Signaling games and stable equilibria. *The Quarterly Journal of Economics* 102(2). 179–221. doi:10.2307/1885060.

Crawford, Vincent P. 2007. Let's talk it over: Coordination via preplay communication with level-k thinking. Unpublished manuscript.

Crawford, Vincent P. & Nagore Iriberri. 2007. Fatal attraction: Salience, naïveté, and sophistication in experimental "hide-and-seek" games. *The American Economic Review* 97(5). 1731–1750. doi:10.1257/aer.97.5.1731.

Farrell, Joseph. 1993. Meaning and credibility in cheap-talk games. *Games and Economic Behavior* 5(4). 514–531. doi:10.1006/game.1993.1029.

Farrell, Joseph & Matthew Rabin. 1996. Cheap talk. *The Journal of Economic Perspectives* 10(3). 103–118. http://www.jstor.org/stable/2138522.

Fine, Kit. 1975. Critical notice: Counterfactuals. *Mind* 84(1). 451–458. doi:10.1093/mind/LXXXIV.1.451.

Fox, Danny. 2007. Free choice and the theory of scalar implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositional semantics*, 71–120. Hampshire: Palgrave Macmillan.

Fox, Danny & Benjamin Spector. 2009. Economy and embedded exhaustification. Handout for a talk given at Cornell University. http://web.mit.edu/linguistics/people/faculty/fox/Fox_Spector_Cornell.pdf.

Franke, Michael. 2009. *Signal to act: Game theory in pragmatics*. Amsterdam: Universiteit van Amsterdam dissertation.

Franke, Michael. 2010. Semantic meaning and pragmatic inference in non-cooperative conversation. In Thomas Icard & Reinhard Muskens (eds.), *Interfaces: Explorations in logic, language and computation* (Lecture Notes in Artificial Intelligence 6211/2010), 13–24. Berlin and Heidelberg: Springer-

Verlag. doi:10.1007/978-3-642-14729-6_2.
Gazdar, Gerald. 1979. *Pragmatics: Implicature, presupposition, and logical form*. New York: Academic Press.
Geurts, Bart. 2005. Entertaining alternatives: Disjunctions as modals. *Natural Language Semantics* 13(4). 383–410. doi:10.1007/s11050-005-2052-4.
Geurts, Bart. 2010. *Quantity implicatures*. Cambridge: Cambridge University Press.
Geurts, Bart & Nausicaa Pouscoulous. 2009a. Embedded implicatures?!? *Semantics & Pragmatics* 2(4). 1–34. doi:10.3765/sp.2.4.
Geurts, Bart & Nausicaa Pouscoulous. 2009b. Free choice for all: A response to Emmanuel Chemla. *Semantics & Pragmatics* 2(5). 1–10. doi:10.3765/sp.2.5.
Grafen, Alan. 1990. Biological signals as handicaps. *Journal of Theoretical Biology* 144(4). 517–546. doi:10.1016/S0022-5193(05)80088-8.
Grice, Paul Herbert. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Speech acts* (Syntax and Semantics 3), 41–58. New York: Academic Press.
Groenendijk, Jeroen & Martin Stokhof. 1984. *Studies in the semantics of questions and the pragmatics of answers*. Amsterdam: Universiteit van Amsterdam dissertation.
Ho, Teck-Hua, Colin Camerer & Keith Weigelt. 1998. Iterated dominance and iterated best response in experimental 'p-beauty contests'. *The American Economic Review* 88(4). 947–969.
Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*. Los Angeles: University of California, Los Angeles dissertation.
Horn, Laurence R. 1984. Towards a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Deborah Shiffrin (ed.), *Meaning, form, and use in context*, 11–42. Washington: Georgetown University Press.
Horn, Laurence R. 1989. *A natural history of negation*. Chicago: Chicago University Press.
Jäger, Gerhard. 2007. The evolution of convex categories. *Linguistics and Philosophy* 30(5). 551–564. doi:10.1007/s10988-008-9024-3.
Jäger, Gerhard & Christian Ebert. 2009. Pragmatic rationalizability. In Arndt Riester & Torgrim Solstad (eds.), *Proceedings of Sinn und Bedeutung 13*, 1–15.
de Jager, Tikitu & Robert van Rooij. 2007. Explaining quantity implicatures. In *Proceedings of the 11th conference on theoretical aspects of rationality and knowledge (TARK)*, 193–202. New York: Association for Computing Machinery. doi:10.1145/1324249.1324276.

Kamp, Hans. 1973. Free choice permission. *Proceedings of the Aristotelian Society* 74. 57–74. http://www.jstor.org/stable/4544849.

Kamp, Hans. 1978. Semantics versus pragmatics. In Franz Guenthner & Siegfried Josef Schmidt (eds.), *Formal semantics and pragmatics for natural languages*, 255–287. Dordrecht: Reidel.

Katzir, Roni. 2007. Structurally-defined alternatives. *Linguistics and Philosophy* 30(6). 669–690. doi:10.1007/s10988-008-9029-y.

Klinedinst, Nathan. 2006. *Plurality and possibility*. Los Angeles: University of California, Los Angeles dissertation.

Kratzer, Angelika. 1981. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10(2). 201–216. doi:10.1007/BF00248849.

Kratzer, Angelika & Junko Shimoyama. 2002. Indeterminate pronouns: The view from Japanese. In Yukio Otsu (ed.), *Proceeding of the 3rd Tokyo conference on psycholinguistics*, 1–25.

van Kuppevelt, Jan. 1996. Inferring from topics: Scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy* 19(4). 393–443. doi:10.1007/BF00630897.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Lewis, David. 1969. *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Lewis, David. 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.

Matsumoto, Yo. 1995. The conversational condition on Horn scales. *Linguistics and Philosophy* 18(1). 21–60. doi:10.1007/BF00984960.

McClure, William. 2000. *Using Japanese—a guide to contemporary usage*. Cambridge: Cambridge University Press.

McKay, Thomas & Peter van Inwagen. 1977. Counterfactuals with disjunctive antecedents. *Philosophical Studies* 31(5). 353–356. doi:10.1007/BF01873862.

Merin, Arthur. 1992. Permission sentences stand in the way of boolean and other lattice-theoretic semantices. *Journal of Semantics* 9(2). 95–162. doi:10.1093/jos/9.2.95.

Mühlenbernd, Roland. 2009. *Kommunikationsmodell für den Entwicklungsprozess von Implikaturen*: University of Bielefeld MA thesis.

Nute, Donald. 1975. Counterfactuals and the similarity of worlds. *Journal of*

Philosophy 72(21). 773–778. doi:10.2307/2025340.
Parikh, Prashant. 1991. Communication and strategic inference. *Linguistics and Philosophy* 14(5). 473–514. doi:10.1007/BF00632595.
Parikh, Prashant. 2001. *The use of language*. Stanford, CA: CSLI Publications.
Rabin, Matthew. 1990. Communication between rational agents. *Journal of Economic Theory* 51(1). 144–170. doi:10.1016/0022-0531(90)90055-O.
van Rooij, Robert. 2000. Permission to change. *Journal of Semantics* 17(2). 119–143. doi:10.1093/jos/17.2.119.
van Rooij, Robert. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy* 26(6). 727–763. doi:10.1023/B:LING.0000004548.98658.8f.
van Rooij, Robert. 2004. Signalling games select horn-strategies. *Linguistics and Philosophy* 27(4). 493–527. doi:10.1023/B:LING.0000024403.88733.3f.
van Rooij, Robert. 2006. Free choice counterfactual donkeys. *Journal of Semantics* 23(4). 383–402. doi:10.1093/jos/ffl004.
van Rooij, Robert. 2008. Games and quantity implicatures. *Journal of Economic Methodology* 15(3). 261–274. doi:10.1080/13501780802321376.
van Rooij, Robert. 2010. Conjunctive interpretation of disjunction. *Semantics & Pragmatics* 3(11). 1–28. doi:10.3765/sp.3.11.
van Rooij, Robert & Katrin Schulz. 2004. Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information* 13(4). 491–519. doi:10.1007/s10849-004-2118-6.
van Rooij, Robert & Katrin Schulz. 2006. Only: Meaning and implicatures. In Maria Aloni, Alistair Butler & Paul Dekker (eds.), *Questions in dynamic semantics* (Current Research in the Semantics/Pragmatics Interface 17), 193–223. Amsterdam and Singapore: Elsevier.
Ross, Alf. 1944. Imperatives and logic. *Philosophy of Science* 11(1). 30–46. doi:10.1086/286823.
Russell, Benjamin. 2006. Against grammatical computation of scalar implicatures. *Journal of Semantics* 23(4). 361–382. doi:10.1093/jos/ffl008.
Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27(3). 367–391. doi:10.1023/B:LING.0000023378.71748.db.
Schulz, Katrin. 2005. A pragmatic solution for the paradox of free choice permission. *Synthese* 147(2). 343–377. doi:10.1007/s11229-005-1353-y.
Schulz, Katrin & Robert van Rooij. 2006. Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy* 29(2). 205–250. doi:10.1007/s10988-005-3760-4.
Simons, Mandy. 2005. Dividing things up: The semantics of *or* and the modal/*or* interaction. *Natural Language Semantics* 13(3). 271–316.

doi:10.1007/s11050-004-2900-7.

Spector, Benjamin. 2006. Scalar implicatures: Exhaustivity and Gricean reasoning. In Maria Aloni, Alistair Butler & Paul Dekker (eds.), *Questions in dynamic semantics*, 229–254. Amsterdam and Singapore: Elsevier.

Spector, Benjamin. 2007. Aspects of the pragmatics of plural morphology: On higher-order implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositional semantics*, 243–281. Basingstoke and New York: Palgrave Macmillan.

Spence, Andrew Michael. 1973. Job market signaling. *Quarterly Journal of Economics* 87(3). 355–374. doi:10.2307/1882010.

Stalnaker, Robert. 1968. A theory of conditionals. In Nicholas Rescher (ed.), *Studies in logical theory*, 98–112. Oxford: Oxford University Press.

Stalnaker, Robert. 1998. Belief revision in games: Forward and backward induction. *Mathematical Social Sciences* 36(1). 31–56. doi:10.1016/S0165-4896(98)00007-9.

Stalnaker, Robert. 2006. Saying and meaning, cheap talk and credibility. In Anton Benz, Gerhard Jäger & Robert van Rooij (eds.), *Game theory and pragmatics*, 83–100. Basingstoke and New York: Palgrave Macmillan.

von Stechow, Arnim & Thomas Ede Zimmermann. 1984. Term answers and contextual change. *Linguistics* 22(1). 3–40. doi:10.1515/ling.1984.22.1.3.

Swanson, Eric. 2010. Structurally defined alternatives and lexicalizations of XOR. *Linguistics and Philosophy* 33(1). 31–36. doi:10.1007/s10988-010-9074-1.

Veltman, Frank. 1985. *Logics for conditionals*. Amsterdam: Universiteit van Amsterdam dissertation.

Warmbrōd, Ken. 1981. Counterfactuals and substitution of equivalent antecedents. *Journal of Philosophical Logic* 10(2). 267–289. doi:10.1007/BF00248853.

von Wright, Georg Henrik. 1951. Deontic logic. *Mind* 60(237). 1–15. doi:10.1093/mind/LX.237.1.

von Wright, Georg Henrik. 1968. *An essay on deontic logic and the theory of action*. Amsterdam: North-Holland Publishing Company.

Zimmermann, Thomas Ede. 2000. Free choice disjunction and epistemic possibility. *Natural Language Semantics* 8(4). 255–290. doi:10.1023/A:1011255819284.

Michael Franke
Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstraße 19
72074 Tübingen
michael.franke@uni-tuebingen.de