Signal to Act

Game Theory in Pragmatics

Michael Franke

Signal to Act

Game Theory in Pragmatics

ILLC Dissertation Series DS-2009-11

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 904
1098 XH Amsterdam
phone: +31-20-525 6051
fax: +31-20-525 5206
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Signal to Act

Game Theory in Pragmatics

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D.C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Agnietenkapel
op dinsdag 15 december 2009, te 10.00 uur

door

Michael Franke

geboren te Würselen, Duitsland.

Promotiecommissie

Promotor:
prof.dr. M.J.B. Stokhof

Co-promotor:
dr. R.A.M. van Rooij

Overige leden:
prof.dr. N. Asher
prof.dr. J.A.G. Groenendijk
prof.dr. P. Hendriks
prof.dr. G. Jäger
prof.dr. F.J.M.M. Veltman
dr. H.W. Zeevat

Faculteit der Geesteswetenschappen

Cover design by Michael Franke.
Printed and bound by Ipskamp Drukkers B.V., Enschede.

ISBN: 978–90–5776–202–4

Contents

Contents · vii

Acknowledgments · xi

Overview · xv

1 · What is Game Theoretic Pragmatics? · 1

1.1 · Gricean Pragmatics · 2

1.1.1 · The Gricean Programme · 4

1.1.2 · Conversational Implicatures · 6

1.1.3 · Brands of Griceanism · 9

1.2 · Game Theory for Gricean Pragmatics · 13

1.2.1 · Static Games & Their Solutions · 13

1.2.2 · Signaling Games in Gricean Pragmatics · 17

1.2.3 · Solving Signaling Games · 24

1.2.4 · Implementing Semantic Meaning · 34

2 · The Iterated Best Response Model · 43

2.1 · Focal Points & Iterated Best Response · 44

2.1.1 · Semantic Meaning as a Focal Point · 46

2.1.2 · Iterated Best Response Reasoning as Pragmatic Inference · 48

2.1.3 · Strategic-Type Models · 52

2.2 · The Vanilla Model · 53

2.2.1 · Strategic Types and the IBR Sequence · 54

2.2.2 · Examples: Scalar & M-Implicatures · 59

2.2.3 · Consistency: Naïve & Sophisticated Updates · 66
2.2.4 · Truth Ceteris Paribus & Skewed Priors · 71
2.3 · Forward Induction · 76
2.3.1 · Trouble-Maker "Some But Not All" · 77
2.3.2 · Forward Induction and Strong Belief in Rationality · 78
2.3.3 · Restrictions on Counterfactual Beliefs · 81
2.4 · Overview and Comparison · 89
2.4.1 · Versions of the IBR Model · 89
2.4.2 · Related Models · 91
2.4.3 · IBR vs. Rationalizability · 96
2.5 · Semantic Meaning and Credibility · 107
2.5.1 · Message Credibility · 108
2.5.2 · Credibility-Based Refinements · 113

3 · Games and Pragmatic Interpretation · 123
3.1 · Game Models Revisited · 124
3.1.1 · Interpretation Games · 124
3.1.2 · Construction of Interpretation Games · 126
3.1.3 · Examples: Multiple Scalar Items · 134
3.2 · Epistemic Lifting of Signaling Games · 138
3.2.1 · The Epistemic Status of Scalar Implicatures · 138
3.2.2 · The Implicatures of Plain Disjunctions · 140
3.2.3 · Lifted Signaling Games · 141
3.2.4 · Examples · 145
3.3 · Free Choice Inferences · 156
3.3.1 · Free Choice from Anti-Exhaustivity · 158
3.3.2 · Anti-Exhaustivity from Iteration · 161
3.3.3 · Simplification of Disjunctive Antecedents · 169
3.4 · Games at the Semantics-Pragmatics Interface · 175

4 · Perspective, Optimality & Acquisition · 181
4.1 · Optimality Theory in Pragmatics · 182
4.1.1 · OT-Systems · 182
4.1.2 · Uni- and Bidirectional Optimality · 184
4.1.3 · Example: M-Implicatures in BiOT · 187
4.1.4 · BiOT as a Model of Pragmatic Interpretation · 189

4.2 · BIOT and Game Theory · 190
4.2.1 · BIOT and Strategic Games · 191
4.2.2 · Critique · 192
4.2.3 · BIOT and Signaling Games · 194
4.3 · An Epistemic Interpretation of Optimality · 196
4.3.1 · Comprehension Lags in Language Acquisition · 196
4.3.2 · Unidirectional Optimality · 199
4.3.3 · Bidirectional Optimality · 203
4.4 · Scalar Implicatures in Language Acquisition · 213
4.4.1 · Overview of Some Recent Studies · 213
4.4.2 · Tolerance vs. Conceptualization · 220

5 · The Pragmatics of Conditionals · 225
5.1 · Meaning and Use of Conditionals · 226
5.1.1 · Semantics for Conditionals · 227
5.1.2 · Kinds of Conditionals · 231
5.2 · Conditional Perfection · 234
5.2.1 · Approaching Perfection · 236
5.2.2 · Two Sources of Perfection · 239
5.2.3 · Rationalizing Indirectness · 249
5.2.4 · Forward Induction under Awareness Dynamics · 252
5.3 · Unconditional Readings · 257
5.3.1 · Biscuit Conditionals · 258
5.3.2 · Unconditionality Beyond Biscuits · 262
5.3.3 · Conditional Independence · 265
5.3.4 · Biscuits in Discourse · 274
5.3.5 · Projection and a Big Fat Lie · 277

6 · Conclusions & Outlook · 279

References · 281

Notation, Symbols & Abbreviations · 299

Samenvatting · 301

Abstract · 303

Acknowledgments

> "Abstract work, if one wishes to do it well, must be allowed to destroy one's humanity; one raises a monument which is at the same time a tomb, in which, voluntarily, one slowly inters oneself."
> (Bertrand Russell, in a letter to Lucy Martin Donnelly, May 23, 1902)
>
> "It's incredibly important to unthink at least once a day for the very preservation of the intellectual life." (Allan Watts, *Intellectual Yoga*)

Writing a dissertation is not always easy. It *is* easy on a good day, but not on a bad day. On a good day, our work is not motivated by the *desire* to fulfil a formal requirement towards higher education, but rather by what is felt as a forbidding *necessity* to find answers to a flock of questions that a peculiar twist of our curiosity has excavated as, let us say, relevant. What keeps us going on a good day is the plain beauty of progress, even if imaginary, and the delight of however mild an insight, even if later to be proved wrong.

On a bad day, however, the compulsion for answers, especially for *good* answers, can turn out quite harmful. This is so because the ultimate authority against which we have to defend the quality of our work is essentially our own proud ideal. The main cause of stress therefore is internal, and cannot easily be shrugged off with anger or laughter: what keeps us awake and working at night is not some third-person executioner, which it would be easy to hate or ridicule for alleviation of the pressure, but it is our own heads to which we are a slave. What keeps us going on a bad day is a mystery to me, but I have some vague understanding of what fueled *my* motors in times of need: supportive supervisors, inspiring colleagues, friends and family.

During my whole stay in Amsterdam, my supervisor Robert van Rooij has been an inexhaustible source of inspiration. I was fortunate to have him appointed as my mentor during the Master of Logic programme and I was lucky

to win him as a supervisor of my master thesis. This dissertation was written as part of his NWO project "The Economics of Language" and I am naturally grateful for the trust he demonstrated by offering this position to me. Over the last six years, Robert has spent innumerable hours reading, commenting and discussing ideas of mine. Especially in the beginning of our collaboration, our relation was remarkably similar to a process of evolution but with clearly divided labor between us: I was producing ideas, and he was weeding out the obvious nonsense and selected the fittest of thoughts which carried hope of mutating into something interesting. At later stages, I profited immensely from his rigorous pragmatism which made me start writing in time and abandon irrelevant daydreams, a.k.a. side projects. Without his impetus, I would perhaps still be gazing at the sky from behind the window pondering what a curious beast language is after all. Thanks Robert!

I am also very happy to have benefited from numerous detailed discussions with Martin Stokhof, who I learned to respect deeply for his well-informed advice on how to fine-tune my arguments, in structure and wording, in the light of many looming philosophical pitfalls. I also remember still a dark and difficult day when I was a master student who approached Martin looking for advice on how to make logic and philosophy of language relevant for me "as a human being." I have forgotten, funnily enough, *how* he did it, but he did fix me in an afternoon's conversation, and I have never had the need for a similar logico-psychiatric session ever after. Thanks Martin!

I would also like to thank my colleague and friend Tikitu de Jager, without whose open ears, critical questioning, conversational temperament and delightful eagerness for distractions the last years would have been a lot more dull, boring and lifeless. It is he who is responsible for me using Emacs, for bringing issues of typography to my attention, and it is he who answered all my LaTeX questions, most of which I would not have had if it had not been for him proselytizing me to care. Tikitu also kindly proofread this dissertation in times of utmost busyness, and pointed out a minor mistake in section 2.2.4 that fortunately has no impact on any issue of relevance. Thanks Tikitu!

Many other people have helped me tremendously with their critical thinking in many hours of conversation. Anton Benz has kindly invited me to the Zentrum für Allgemeine Sprachwissenschaft in Berlin once in 2007 and once in early 2009. His theory of "optimal assertions" was of central importance to the development of the model that is presented in this thesis, and I would like to thank Anton for his renewed hospitality, many insightful discussions and general inspiration and support. I also greatly benefited from reading some of

Gerhard Jäger's recent work, that turned out so closely related to mine at the time, it was frightening. Gerhard kindly invited me to Bielefeld and Tübingen where I was able to present late material of this thesis. Needless to say, I have learned a lot from discussions at these occasions.

But also in Amsterdam, many people have had open ears, commented on my work and helped with insightful suggestions. Here, I would like to thank especially Maria Aloni, Reinhard Blutner, Paul Dekker, Floris Roelofson, Katrin Schulz, Frank Veltman and Henk Zeevat. Merging the academic and personal, I have had many very dear and deep conversations with Sven Lauer, Magdalena Schwager and Marc Staudacher. The list of friendly colleagues who kindly helped some way or other in writing this thesis is long, and even at risk of forgetfulness, let me try to mention some of you folks out there: Judith Degen, Cornelia and Christian Ebert, Hannah Gieseler, Nikos Green, Jacqueline Griego, Yurie Hara, Napolean Katsos, Nathan Klinedinst, Fabienne Martin, Eric McReady, Chris Potts, Daniel Rothschild, Tatjana Scheffler and Matthew Wampler-Doty.

I would also like to send out a hearty "jai" to the crowd at Svaha Yoga. I am deeply grateful to Patrick and Gösta for having realized this oasis of calm: had I not regularly cleared my head on the mat, I am tempted to say in exaggeration, I would have lost it at some point.

It is not, however, so easy to exaggerate the admiration I find for the patience and the tolerance of Julie Verburg, who has miraculously accepted my work compulsion as a part of me and made room in her heart for even that. Unbelievable!

Nicht zuletzt, sondern an exponierter Stelle gehöhrt mein Dank auch meinen Eltern ausgesprochen, die in ihrer über Jahrzente augeübten Überzeugung Recht behalten haben, dass ich glücklich werden würde, wenn man mir alle Freiheit schüfe, meinen eigenen Weg zu finden. Die hier vorliegende Arbeit ist eine Teamleistung. Danke für die bedingungslose Rückendeckung.

Overview

This thesis is interdisciplinary in nature. Its main contribution is an application of game theory to linguistic pragmatics. Since perhaps not many people will be familiar with both subjects at once, the need arises to introduce the basics of both fields. Although admittedly the thesis spends more effort on explaining the relevant concepts of game theory to the linguist than on explaining the relevant concepts of pragmatics to the game theorist, I would sincerely hope that the text is accessible, at least in its gist, to anybody proficient in some adjacent academic field who is interested in the topic. Be that as it may, it would certainly be forlorn optimism to expect that all of my possible readers are equally interested in all issues addressed here. I would therefore like to give a brief overview of the content of this thesis, together with an indication which parts belong to either the *linguist's track* or the *game theorist's track*. The linguist's track contains all linguistic applications and only the absolutely necessary information on game theory. The game theorist's track, on the other hand, contains the game theoretic details and only the absolutely necessary information on linguistic pragmatics.

There are five main chapters. Chapter 1 introduces the basics of both Gricean pragmatics and game theory. Chapter 2 spells out the central iterated best response (IBR) model of pragmatic reasoning. Chapter 3 is dedicated to linguistic applications of the IBR model. Chapter 4 compares the IBR model to bidirectional optimality theory and discusses data from language acquisition. Finally, chapter 5 is mainly linguistic and deals with use and interpretation of conditionals. (A more thorough abstract of the thesis can be found at the end, on page 303.)

The linguist's track obligatorily contains sections 1.2, where basic concepts

of game theory are introduced, as well as sections 2.1 and 2.2 to understand the basic IBR model. Reading section 2.3, which discusses a refinement of the IBR model, is also recommended. After that any part or portion of chapters 3, 4 and 5 that seems relevant to the reader's concern should be intelligible. In fact, chapter 5 is nearly independent of the game theoretic framework (with the exception of sections 5.2.4 and 5.3.4).

The game theorist's track obligatorily contains section 1.1, which introduces the basic ideas of Gricean pragmatics, as well the whole of chapter 2. Here especially sections 2.4 and 2.5 are relevant, which compare the IBR model to related game theoretic approaches. The game theorist might furthermore take interest in section 3.1 where I discuss my preferred interpretation of signaling games in a linguistic context. Finally, a cursory glance at some of the applications in sections 3.2 and 3.3 will help understand better the linguistic motivation behind the present approach.

Chapter 1

What is Game Theoretic Pragmatics?

> "To anyone who knew, for instance, my old scout at Oxford, or a certain one of the shopkeepers in the village where I live, it would be ludicrous to suggest that *as a general principle* people's speech is governed by maxims such as 'be relevant'; 'do not say that for which you lack adequate evidence' (!); 'avoid obscurity of expression, ambiguity or unnecessary prolixity' (!!). In the case of the particular speakers I am thinking of (and I have no doubt that any reader could supply his own counterparts), the converse of Grice's maxims might actually have greater predictive power." (Sampson 1982, p. 203)

> "Making sense of the utterances and behavior of others, even their most aberrant behavior, requires us to find a great deal of reason and truth in them." (Davidson 1974, p. 321)

Chapter Contents

1.1 · Gricean Pragmatics · 2

1.2 · Game Theory for Gricean Pragmatics · 13

It is a near-platitude that under normal circumstances we reliably learn more from observing the honest *utterance* of a declarative sentence[1] than we would learn from the direct observation of infallible evidence that the proposition expressed by that sentence was true. If John stands by the window and says

(1) It's raining.

we learn more from his utterance than what we would learn from a glimpse of the wet street outside (assuming for the sake of argument that this counts as infallible evidence for rain). Of course, if John is honest and reliable, we do learn that it is raining from his utterance, just as we would from observation. But depending on the concrete circumstances, John's utterance, but certainly not the observation of the wet street outside, might also inform us that

(2) a. John advises we should take an umbrella, or that
    b. John (hereby) declares the picnic cancelled, or that
    c. John is sick of living in Amsterdam.

These are non-trivial pieces of information that a proficient interpreter gets to understand that go way beyond the meaning of the sentence "It's raining." So where does this information come from? Why is such surplus information reliably inferred and communicated? What role does the conventional, semantic meaning of an utterance play in the process of fully understanding it? What features of the context of an utterance are important for its interpretation? These are the kind of questions that LINGUISTIC PRAGMATICS tries to raise, sharpen and answer.

1.1 Gricean Pragmatics

One way of approaching the difference between utterance and observation is to see an utterance clearly as an instance of human *action*, and as such to subject it to commonsense conceptualization in terms of the speaker's beliefs, preferences and intentions. From this point of view, we may conceive of linguistic pragmatics as an investigation into the systematic relationship between

1. Although declarative sentences usually receive most attention, similar remarks could be made about non-declarative sentences, phrases, words, gestures or any other kind of ostensive behavior with a sufficient history of preceding uses to bestow an element of commonly expected meaningfulness to it.

the conventional, semantic meaning of a linguistic token and the overall significance that it may acquire when put to use in human action in a concrete context.[2]

It clearly has a certain appeal to distinguish aspects of meaning that belong to the meaningful sign proper and those that arise from the reasons and ends for which a meaningful sign is used. For instance, we would not want to hold that the sentence (1) itself contains ambiguously all the possible further shades of meaning it might acquire in special contexts. This is because the list of such special contextualized meanings would be enormous if not infinite. A mere list of possible situated meanings would moreover be less explanatory than one could possibly hope for, because it might conceal certain regularities in the interaction of conventional meaning and contextual use, so much so as to possibly even undermine any reasonable concept of semantic meaning.

This view is clearly corroborated by inferences that appear rather rule-like — inferences that are tied closely, for instance, to the use of a particular lexical item. A standard example here is the quantifier phrase "some." In most situations an utterance of the sentence (3a), may reliably convey the inference in (3b).[3]

(3) a. I saw some of your children today.

b. The speaker did not see *all* of the hearer's children today.

But would we want to say that "some" semantically means "some and not all"? Preferably not, many philosophers of language have argued, because, among other things, the attested inference can be easily *cancelled* as in (4).

(4) I saw some of your children today, and maybe even all of them.

2. This view of pragmatics still resembles the distinction of semiotic subdisciplines into syntax, semantics and pragmatics which was introduced by Charles M. Morris: while syntax studies the relation between signs, and semantics the relation between signs and objects, pragmatics "deals with the origins, uses, and effects of signs within the total behavior of the interpreters of signs" (Morris 1946, p. 219).

3. To be precise, the inference that sentence (3a) gives rise to has either a stronger or a weaker epistemic reading (Gazdar 1979; Soames 1982):

(1) The speaker knows/believes that she did not see all of the hearer's children.

(2) The speaker does not know/believe that she saw all of the hearer's children.

I will come back to this issue only very late in this thesis, namely in chapter 3 which deals extensively with linguistic applications and inferences about the speaker's doxastic state.

We should also not assume that "some" is lexically ambiguous, because the phenomenon lends itself to a much more interesting and systematic explanation. This argument has already been advanced by John Stuart Mill in the 19th century in a response to an ambiguity thesis proposed by William Hamilton:

> "No shadow of justification is shown (...) for adopting into logic a mere sous-entendu of common conversation in its most unprecise form. If I say to any one, 'I saw some of your children to-day', he might be justified in inferring that I did not see them all, not because the words mean it, but because, if I had seen them all, it is most likely that I should have said so: even though this cannot be presumed unless it is presupposed that I must have known whether the children I saw were all or not."
>
> (Mill 1867)

1.1.1 The Gricean Programme

Roughly a century later, Herbert Paul Grice reiterated Mill's position in his William James Lectures, presented at Harvard in 1967. In a condensed formulation that has become known as Grice's *Modified Occam's Razor* he demanded that "senses are not to be multiplied beyond necessity" (Grice 1989, p. 47).[4] Grice's main contribution to a defense of parsimony in logical semantics was the proof that the pragmatic inferences in question can be explained *systematically* based on certain assumptions about proper conduct of a conversation. Grice hypothesized that in most normal circumstances interlocutors share a common core of convictions about the purpose of a conversation and behave, in a sense, *rationally* towards this commonly shared end. This regularity in linguistic behavior explains, so Grice's conjecture, pragmatic inferences of the attested sort.

Maxims of Conversation. In particular, Grice proposed to view conversation as guided by an overarching Cooperative Principle, formulated as a rule of conduct for speakers:

> Cooperative Principle: "Make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction

4. The name of Grice's postulate is chosen in reference to 'Occam's Razor' a principle loosely attributed to the 14th century philosopher William of Occam (though apparently not found in his writing), which pleads for ontological parsimony in theorizing: "Entia non sunt multiplicanda praeter necessitatem."

> of the talk exchange in which you are engaged."
>
> (Grice 1989, p. 26)

Subordinated to the Cooperative Principle, Grice famously gave a perspicuous set of guidelines for proper speaker conduct in his Maxims of Conversation:

> Maxim of Quality: Try to make your contribution one that is true.
>
> (i) Do not say what you believe to be false.
>
> (ii) Do not say that for which you lack adequate evidence.
>
> Maxim of Quantity:
>
> (i) Make your contribution as informative as is required for the current purposes of the exchange.
>
> (ii) Do not make your contribution more informative than is required.
>
> Maxim of Relation:
>
> (i) Be relevant.
>
> Maxim of Manner: Be perspicuous.
>
> (i) Avoid obscurity of expression.
>
> (ii) Avoid ambiguity.
>
> (iii) Be brief (avoid unnecessary prolixity).
>
> (iv) Be orderly. (Grice 1989, pp. 26–27)

Grice showed that hearers can reliably and systematically interpret utterances and infer additional information that goes beyond the semantic meaning of the uttered sentence, based on the assumption that the speaker obeys the Cooperative Principle and the Maxims of Conversation. The main idea of the Gricean Programme is thus to make pragmatic inference amenable to systematic investigation, and to find regularities and structure in conversational behavior and natural language interpretation. Indeed, this idea has had tremendous impact on the philosophy of language and linguistic pragmatics, inspiring and spawning a whole industry of literature on topics and problems raised by Grice's work.[5]

5. For more on the impact of Grice's work see Neale (1992) and Chapman (2005).

1.1.2 Conversational Implicatures

In order to separate aspects of meaning that belong to a conventional sign proper and those that arise from aspects of its use, Grice coined the term of art IMPLICATURE (see Levinson 1983; Horn 2004, for general overview). Being obviously very aware of many looming problems, Grice himself eschewed a proper definition, but on rough approximation it is in his spirit to say that an implicature of an utterance is an aspect of what was *meant* by an utterance but not (literally) *said*.

Some implicatures Grice called CONVENTIONAL IMPLICATURES in the sense that they are associated —as it were by convention— with certain lexical items or specific syntactic constructions (Karttunen and Peters 1974; Bach 1999; Potts 2005). A common example of a conventional implicature is the English sentential connective "but" as in (5) which communicates some adversary relation or contrast between conjuncts on top of logical conjunction.[6]

(5) a. Aino is young but outstandingly clever.

b. $\rightsquigarrow$ Since Aino is young, *it is unexpected* that she is so clever.

From conventional implicatures, Grice distinguished CONVERSATIONAL IMPLICATURES. What crucially sets these two kinds of implicatures apart is that the latter are CALCULABLE in a sense that the former are not: Grice held that it is a defining mark of conversational implicatures that they can be reconstructed as an inference. In the words of Grice himself:

> "The presence of a conversational implicature must be capable of being worked out; for even if it can in fact be intuitively grasped, unless the intuition is replaceable by an argument, the implicature (if present at all) will not count as a conversational implicature; it will be a conventional implicature." (Grice 1989, p. 31)

More in particular, Grice considered conversational implicatures as aspects of meaning that can be backed up or justified by a reasoning process that takes into account the semantic meaning of the utterance, as well as certain aspects of the conversational context. Furthermore, the inference by which a conversational implicature can be derived would in some fashion involve the Cooperative Principle and the Maxims of Conversation: a conversational

6. I use the symbol $\rightsquigarrow$ to mark a possible candidate implicature that an utterance of a given sentence has or might have in a standard context of its use.

implicature is either a direct consequence of the speaker obeying the conversational postulates, or it arises from the speaker's obvious and ostensible opting out of or flouting the maxims.

> "To work out that a particular conversational implicature is present, the hearer will rely on the following data: (1) the conventional meaning of the words used, together with the identity of any references that may be involved; (2) the Cooperative Principle and its maxims; (3) the context, linguistic or otherwise, of the utterance; (4) other items of background knowledge; and (5) the fact (or supposed fact) that all relevant items falling under the previous headings are available to both participants and both participants know or assume this to be the case."
>
> (Grice 1989, p. 31)

Scalar Implicatures. The most prominent examples of conversational implicatures are scalar implicatures. The above example (3) is an instance thereof which hinges on the comparison of scalar expressions "some" and "all." Other examples are the following:

(6) a. It's possible that Yuuki is coming late again.
    b. $\rightsquigarrow$ It's not certain/necessary that Yuuki is coming late again.

(7) a. Hanako sometimes listens to jazz.
    b. $\rightsquigarrow$ Hanako does not often/always listen to jazz.

The abstract reasoning pattern behind a scalar inference seems to be the following naïve scalar reasoning:[7] an utterance of a sentence $S[X]$ which contains a scalar expression X needs to be compared to other possible utterances, in particular to utterances of sentences $S[X']$ where X is replaced with an alternative expression $X' \in \text{Alt}(X)$ from a set of reasonable alternatives to X; an utterance of $S[X]$ then conveys the scalar implicature that all those sentences $S[X']$ are not true which are more informative in virtue of their semantic meaning, i.e., which semantically entail $S[X]$, and whose extra information would have been relevant for the shared cooperative purpose of the conversation. This inference pattern is clearly a sharpened application of especially Grice's Maxim of Quantity.

7. This formulation does not aim to be faithful to any particular proposal, but rather aims at distilling, in rough approximation, the common and very intuitive core idea behind a variety of approaches to scalar reasoning (cf. Horn 1972; Gazdar 1979; Levinson 1983; Horn 1984).

For instance, an utterance of the sentence in (8a) with the scalar expression "some" would be compared to possible utterances of sentences in (8b)–(8d) based on a set of *lexical alternatives* for "some" such as:

$$\text{Alt}(\text{some}) = \{\text{few},\ \text{most},\ \text{all}\}\,.$$

Since (8c) and (8d) semantically entail (8a), we derive the implicatures in (8f) and (8g); but since (8b) does not semantically entail (8a), the implicature in (8e) is not derived by the naïve scalar reasoning pattern.

(8) a. Some of Kiki's friends are metalheads.
b. Few of Kiki's friends are metalheads.
c. Most of Kiki's friends are metalheads.
d. All of Kiki's friends are metalheads.
e. $\not\rightsquigarrow$ It's not the case that few of Kiki's friends are metalheads.
f. $\rightsquigarrow$ It's not the case that most of Kiki's friends are metalheads.
g. $\rightsquigarrow$ It's not the case that all of Kiki's friends are metalheads.

Horn's Division of Pragmatic Labor. Another fairly systematic pattern of pragmatic inference is what has become known as Horn's division of pragmatic labor. It is a fairly ubiquitous phenomenon in natural languages that a simple way of expressing a meaning (9a) is associated with a stereotypical interpretation (9b), whereas a marked and overly complex way of expressing the very same meaning (10a) is interpreted in a non-stereotypical way (10b).

(9) a. Black Bart killed the sheriff.
b. $\rightsquigarrow$ Black Bart killed the sheriff in a stereotypical way.

(10) a. Black Bart caused the sheriff to die.
b. $\rightsquigarrow$ Black Bart killed the sheriff in a non-stereotypical way.

On closer look, Horn's division of pragmatic labor actually captures the interplay of two inferences. In abstract terms, there are two semantically equivalent expressions m and m' both of which could denote either an unmarked t, or a marked state of affairs t'. Given that one expression m' is more marked than the other m, the first part of the pragmatic inference pattern associates the unmarked form with the unmarked state of affairs ($m \rightsquigarrow t$); the second part of the pragmatic inference pattern associates the marked form with the marked state of affairs ($m' \rightsquigarrow t'$).

This double inference plausibly revolves around the Maxim of Quantity and possibly also the Maxim of Manner. Horn (1984) originally described the pattern as arising from the interplay of the two submaxims of Quantity.[8] Levinson (2000) stressed the role of the Maxim of Manner and introduced a further M-principle with which to explain this inference, which is why we could also speak of M-IMPLICATURES here.[9] In the following, I will specifically use this term to refer to the second part of the inference pattern, the association $m' \rightsquigarrow t'$ of marked expressions with marked meanings, without necessarily endorsing Levinson's theory.

Scalar implicatures and M-implicatures will accompany us through the rest of this chapter, as well as the following, as running examples for many of the concepts and notions we will encounter.

1.1.3 Brands of Griceanism

To say that Grice's contribution was heavily influential is not to imply that it was entirely uncontroversial. Even to those who wholeheartedly embarked on the Gricean Programme the exact formulation of the maxims seemed a point worth improvement. It was felt that —to say it with a slightly self-referential twist— the Gricean maxims did not do justice to themselves, in particular to the Maxim of Manner, being long-winded and too vague to yield precise predictions in a number of linguistically relevant cases. Over the years, many attempts have been made to *refine* and *reduce* the Gricean maxims.

NEO-GRICEAN PRAGMATICS. A particularly prominent and successful strand of maxim reduction is found in the work of so-called NEO-GRICEANS (Horn 1972; Gazdar 1979; Atlas and Levinson 1981; Levinson 1983; Horn 1984). This work is largely in keeping with the Gricean assumption of cooperation in conversation and seeks to explain pragmatic inference mostly by a refined explication of the Maxim of Quantity, thereby placing the main emphasis on the role of informativity in discourse. The Neo-Griceans recast the Maxim of

8. More specifically, Horn derived the inference from the interaction of the Q- and I-principle as requirements on speaker and hearer economy (see below). This then also explains the label 'division of pragmatic labor.'

9. Levinson's M-principle requires speakers to use marked expressions for marked meanings, thus directly hard-wiring half of the to-be-explained inference pattern in a conversational postulate (see Levinson 2000, pp. 135–153).

Quantity as consisting of two interdependent principles, called Q-PRINCIPLE and I-PRINCIPLE (see in particular Horn 1984; Levinson 2000):

(11) Q-PRINCIPLE Say as much as you can (given I).

I-PRINCIPLE Say no more than you must (given Q).

These principles are essentially opposing constraints on the organization of discourse, where the Q-Principle aims to capture the hearer's interest in specificity of information, so as to minimize his efforts in arriving at the correct interpretation, while the I-Principle aims to capture the speaker's interest in efficient language use, so as to minimize her efforts in encoding meaning and producing linguistic forms.

Implicatures derived primarily from either of these principles have been called Q-IMPLICATURES and I-IMPLICATURES respectively: Q-implicatures are synonymous with scalar implicatures; I-implicatures are inferences to stereotype such as:

(12) John has a very efficient secretary.
$\rightsquigarrow$ John has a very efficient *female* secretary.

On top of a systematic classification of conversational implicatures, the Neo-Griceans particularly added tractability to Gricean pragmatics by formally spelling out the reasoning process by which implicatures would be established (see especially Gazdar 1979). It is the Neo-Gricean's ideal of formal clarity of definition and predictions that the present study seeks to maintain and occasionally improve on.

RELEVANCE THEORY. Another prosperous school of research that arose from a critique of Grice's maxims is RELEVANCE THEORY (Sperber and Wilson 1995, 2004), according to which the Maxim of Relation deserves the main role in a theory of interpretation. Crucially, relevance theory explicitly sees itself as a *cognitive theory*, rather than a mere addition to a logico-semantic account of meaning, and we may say that, in this and other respects, relevance theory is less Gricean than, for instance, the Neo-Griceans. Relevance theorists sometimes refer to their position as POST-GRICEAN, clearly indicating that relevance theory abandons the Cooperative Principle and leaves behind the Maxims of Conversation in favor of an interpretation principle framed in terms of cognitive effects and processing efforts.

Though some of its proponents may consider it a strength of relevance theory that its basic notions and operations are not backed up by mathematical formalism, I consider this a regrettable weakness of the theory. Relevance theory seems to trade in the ideal of precision and perspicuity in definition and prediction for another noble virtue: appeal to cognitive plausibility, and more recently also endorsement of empirical data (see Noveck and Sperber 2004). Following relevance theory in this latter respect, but not in the former, the theory of pragmatic interpretation featured in this thesis also subscribes to the ideal of psychological plausibility, both introspectively and empirically.

Optimality Theory. Optimality theoretic pragmatics is another, more formal, approach to Gricean pragmatics which originally built on Neo-Gricean approaches (Blutner 1998, 2000; Blutner and Zeevat 2008). Just like the latter, optimality theoretic pragmatics distinguishes clearly a speaker and a hearer perspective in economizing effort in production and comprehension. The competition of these forces results in various notions of optimality for either production alone, comprehension alone, or both at the same time. Optimality theory then explicitly focuses on issues of *perspective taking* in language use: speakers need to take the hearer's interpretation behavior into account, while hearers need to take the speaker's production behavior into account. The model presented in this thesis also shows a strong appeal to issues of perspective taking (so much so that chapter 4 is dedicated entirely to a thorough investigation of this matter by a direct comparison of optimality theoretic with game theoretic models of pragmatic interpretation).

Gricean Pragmatics and Rationality. While Neo-Griceans foreground the Maxim of Quantity in natural language interpretation, and while relevance theorists emphasize the role of a cognitively informed notion of communicative relevance, Grice himself held that the grounds for his communicative principles were to be found in human *rationality*. He wrote:

> "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 1989, p. 28)

And, also:

> "I am enough of a rationalist to want to find a basis that underlies these facts [i.e. the way people in fact communicate], undeniable though they may be; I would like to be able to think of the standard type of conversational practice not merely as something that all or most do *in fact* follow but as something that it is *reasonable* for us to follow, that we *should not* abandon." (Grice 1989, p. 29)

Picking up Grice's conjecture about a rational foundation of his maxims, early work of Kasher (1976) sought to deduce Grice's maxims from a single postulate of human rationality in action. Many others have since taken this idea further, by giving derivations of Gricean maxims, or similar Grice-inspired postulates, also in game theoretical terms (e.g. Hintikka 1986; Parikh 1991; Asher et al. 2001; van Rooij 2003*a*; de Jager and van Rooij 2007; Rothschild 2008). For linguistic pragmatics, however, the question is not so much whether Grice's *maxims* can be reduced to rationality, but rather whether the *data*, i.e., the particular production and interpretation behavior we would like to explain in terms of the maxims, can be explained well as rational behavior. Consequently, this thesis will *not* be concerned with scrutinizing, rationalizing or even just discussing the Gricean maxims; the maxims and their particular formulation will *not* play any noteworthy role in this thesis. Rather, this thesis will offer models of language use —production and comprehension— in which conversationalists' mutually assumed rationality will be a driving explanatory element.

Game Theoretic Pragmatics. This is where a formal theory of rational human agency in the form of game theory enters. Game theory offers mathematical models of interactive decision making of (mostly: idealized and rational) agents. Game theoretic pragmatics (GTP), as conceived of in this thesis, seeks to apply these models and methods of theoretical economics to Gricean pragmatics. Eventually, this thesis will present a general model of natural language use and interpretation as an application of game theory. Pragmatic competence is to be modelled in the abstract as behavior of idealized agents in a game situation.

Obviously, the appeal to abstract mathematical models in GTP is to pay respect to the ideal of maximal clarity in a pragmatic theory. As such this thesis is of course not the first text to appeal to decision theoretic or game theoretic concepts. Game theory has been applied to the study of implicatures in many forms, be that from a rationalistic perspective (e.g Parikh 1991; Benz and van Rooij 2007), or from a more diachronic point of view (e.g van Rooij 2004*b*;

Jäger 2007).[10] Still, the present endeavour is inspired especially by the hope that the Gricean Programme can be carried further in game theoretic terms by adding empirically supported assumptions about particular features and limitations of the cognitive architecture of reasoners. This is why the present study focuses much more strongly on an EPISTEMIC APPROACH to game theory: by making explicit the role of belief formation and reasoning in an abstract interactive situation we can reasonably implement empirically attested and introspectively plausible assumptions about the psychology of reasoners in general and language users in particular. So let us first lay the foundation for such an approach by introducing the necessary concepts of game theory with due emphasis on their respective epistemic interpretation.

1.2 Game Theory for Gricean Pragmatics

Game theory is a branch of applied mathematics that seeks to model human decision making in complex interactive situations.[11] A GAME in its technical sense is a mathematical structure that abstractly represents a decision situation of several agents, where the outcome of the decisions of each agent depends on the choices of the other agents. For what follows it is important to understand that games, in the technical sense of the word, are *not* models of interactive reasoning or decision making, but only of the situations in which agents engage in this kind of deliberation and choice. It is not the game but a SOLUTION CONCEPT that describes —or, depending on the preferred interpretation, *pre*scribes— actual reasoning and/or decision making. An example of a simple game with some of its possible solution concepts will make this distinction clear.

1.2.1 Static Games & Their Solutions

PRISONER'S DILEMMA. A well-known idealized example situation that game theory models is the so-called PRISONER'S DILEMMA. The prisoner's dilemma is a situation where two individuals are charged with a crime and are held imprisoned with no chance to communicate. Both of the accused are forced

10. For further general assessments of applications of game theory in linguistics see for instance Parikh (2001), Benz et al. (2006), van Rooij (2006*b*) and Jäger (2008*a*).

11. For general introductions to game theory see Myerson (1991); Gibbons (1992); Osborne and Rubinstein (1994); Osborne (2004). A good survey of game theory in a linguistic context is the introduction to Benz et al. (2006).

to either confess the crime or deny it. Both agents know (and know that both know...) that the jury will adjudge the following sentences, depending on whether the accused confess or deny: if only one of them confesses, she who confessed will be sent to jail for a long period, say 10 years, while she who denied will be released. If both the accused confess, they will both go to jail for only a short period, say 2 years. But if both the accused deny the crime, they will both go to jail for an intermediate period of, e.g., 5 years. Clearly, in this situation the outcome of each individual decision depends on the decision of the other, and we can model the case as a game. — What kind of game?

KINDS OF GAMES. Game theory distinguishes different kinds of games, traditionally classified along two dimension: (i) whether the agents' choices are *simultaneous or in sequence*, and (ii) whether all agents have *complete or incomplete information*. Games where players move simultaneously are called STATIC GAMES (alt.: strategic games); games where players move in sequence are called DYNAMIC GAMES (alt.: sequential games). We say that a player has COMPLETE INFORMATION in a game if she knows all the decision relevant details except for the play of other players. In game theoretic jargon, a player who knows the action choices of her opponents has PERFECT INFORMATION. Standardly, game theory assumes players to be imperfectly informed, so that individual decision making crucially depends on conjectures about other players' behavior. It is in this sense that games model decisions in interactive situations.

STRATEGIC GAMES OF COMPLETE INFORMATION. Obviously then, the game that models the prisoner's dilemma is a *strategic game with complete information*, because both accused must make their decision simultaneously (at least in the sense that neither will come to know the decision of the other before she has to make her own decision) and the potential outcome of each combination of simultaneous individual choices is common knowledge between the two. Formally, a STRATEGIC GAME WITH COMPLETE INFORMATION is a triple

$$\langle N, (A)_{i \in N}, (\succeq)_{i \in N} \rangle$$

where N is a set of players, A_i are the actions available to player i and $\succeq_i$ is player i's preference relation over possible outcomes of the game, represented here in terms of ACTION PROFILES $\times_{j \in N} A_j$, i.e., tuples of all possible combinations of individual choices.

	c	d
c	2,2	0,3
d	3,0	1,1

Figure 1.1: The prisoner's dilemma

This structure captures the essential parts of the prisoner's dilemma, for instance as in figure 1.1: this table shows the action choices of two players, one of which —the ROW PLAYER— chooses from the actions c (confess) and d (deny) in the rows of the table, and one of which —the COLUMN PLAYER— chooses from the actions c (confess) and d (deny) in the columns of the table. Player i's preference relation $\succeq_i$ is given in terms of numerical PAYOFFS that are listed as pairs of numbers, one pair for each action profile, where conventionally the row player's payoffs are given first and higher numbers represent individually more preferable outcomes. So, for example, the row player prefers an outcome $\langle c, c\rangle$ where her payoff is 2 to an outcome $\langle c, d\rangle$ where her payoff is zero. As the reader will be able to quickly verify, the payoffs listed in figure 1.1 are consistent with the natural assumption that both agents are interested only in minimizing the duration of their own imprisonment.

NASH EQUILIBRIUM. This game is really just a model of the situation and does not specify what the agents in fact do, or what they should do if they are rational (or, even, what they should not do if they care for each other, for instance). The well-known notion of Nash equilibrium is one possible *solution concept* for this game which specifies the idealized behavior of agents in the situation that is modelled by the game. Formally, a NASH EQUILIBRIUM of a strategic game is an action profile a^* such that for all $i \in N$ there is no $a_i \in A_i$ for which:

$$(a^*_{-i}, a_i) \succ_i a^*.$$

Here (a^*_{-i}, a_i) is the tuple that results from replacing the i-th component of a^* with a_i. In words, a Nash equilibrium is a set of action choices, one for each player, which no single player has an incentive to deviating from, given that all the other players conform. For instance, the only Nash equilibrium of the prisoner's dilemma in figure 1.1 is the action profile $\langle d, d\rangle$ where both agents deny the crime.

Interpretation of Equilibrium. The most common interpretation of Nash equilibrium is as a *steady state* in the behavior of agents when repeatedly playing the game (cf. Heap and Varoufakis 2004; Osborne 2004). As such, the notion does not actually spell out the reasoning process by which a player may arrive at the conclusion that she should —in whatever sense of the modal— be playing a Nash choice, i.e., her part of a Nash equilibrium. Players could arrive at playing their unique Nash choice d in the prisoner's dilemma by a process of gradual improvement over time. In this way it is totally conceivable for players to simply gradually habituate themselves into their (coordinated) Nash choices by small steps of unsophisticated diachronic optimization, such as modelled in *evolutionary game theory* (see Weibull 1997), or by however more sophisticated mechanisms of *learning* (see Fudenberg and Levine 1998).

Thus conceived, Nash equilibrium as a solution concept for strategic games does not —contrary to a seemingly widespread misconception— crucially appeal to a notion of rationality in a player's reasoning about the behavior and beliefs of other players. Even though for a given Nash equilibrium each player's Nash choice is —in an intuitive sense— a rational and optimal response to what everybody else is doing, it is not necessary for Nash equilibrium that any player actually believes that any other player is rational (see Stalnaker 1994; Aumann and Brandenburger 1995).

Rationalizability & Common Belief in Rationality. There are other solution concepts that are more explicitly linked to the reasoning process of agents in one-shot strategic situations, where a game is played only once and for the very first time. One such is rationalizability and its corresponding algorithm called iterated strict dominance (Bernheim 1984; Pearce 1984). The idea behind iterated strict dominance is fairly simple: when confronted with a game such as the prisoner's dilemma in figure 1.1, an agent may reason to herself that playing action c is never an optimal choice *no matter what* the opponent is doing. Action c is strictly dominated by action d in this game, because for all conceivable moves of the opponent choosing d instead of c guarantees a strictly better outcome. Removing all strictly dominated actions from a game may render further actions of either player strictly dominated. Iterating this strictly eliminative process further we will end up with a set of rationalizable actions: all those actions for player i which are no longer strictly dominated by any remaining actions for player i given the remaining actions of player j. For example, in the prisoner's dilemma only action d is rationalizable for both players.

The set of rationalizable actions deserves this name because the algorithmic iteration procedure that leads to it has a straightforward interpretation in terms of an agent's beliefs about rationality. If an action is strictly dominated, it would simply be irrational to choose it. But that means that if player i believes in the rationality of player j, player i will come to believe that player j will not choose a strictly dominated action. Iterated elimination of strictly dominated strategies then corresponds to deeper nestings of belief in rationality. The rationalizable actions are all those actions that are rational choices under common belief in rationality. Hence, unlike equilibrium, rationalizability offers a compelling argument centered on the concept of rationality which leads agents to proper choices in a game situation by mere introspection and deliberation.

Game Models & Solutions. It transpires that any application of game theory generally requires to decide on (i) a proper game model, and (ii) an adequate solution concept. Both choices need to be well motivated individually. The game model should capture all (and, preferably only those) contingencies that are relevant for the phenomenon we would like to describe or explain. Similarly, the solution concept should also be conceptually adequate in the sense that its *epistemic characterization* in terms of agents' beliefs, their reasoning strategies and cognitive capabilities, should fit the overall descriptive or explanatory purpose. Thus, the obvious question for game theoretic approaches to pragmatics becomes: which game model and which solution concept should we consult when exercising ourselves in Gricean pragmatics?

1.2.2 Signaling Games in Gricean Pragmatics

Although static games with complete information, such as models for the prisoner's dilemma, are the easiest and most manageable kinds of games, they are unfortunately not the most natural choice for a model of language use and interpretation. Utterances and their pragmatic reception are rather to be modelled as *dynamic* games, because we would like to capture the sequential nature of utterance and subsequent reception/reaction and the natural asymmetry in information between interlocutors. Of course, different kinds of utterances would require different kinds of dynamic games. For instance, in modelling a run-of-the-mill case of an informative assertion the speaker should possess information that the hearer lacks, whereas in the case of a stereotypical information-seeking question we would like to look at a game

in which the speaker is uninformed while the hearer is potentially informed.

In general, I suggest that a (dynamic) game should be regarded as a reduced and idealized, but for certain purposes sufficient, MODEL OF THE UTTERANCE CONTEXT: it represents a few (allegedly: *the most*) relevant parameters of a conversational context, viz., the interlocutors' beliefs, behavioral possibilities and preferences, in rather crude, idealized abstraction. This general, conceptual point will become clearer when we look at an easy example of a dynamic game and its interpretation as a context model.

THE WINE-CHOICE SCENARIO. Suppose Alice is preparing dinner for her visitor Bob who would like to bring a bottle of wine. Depending on whether Alice prepares beef or fish, Bob would like to bring red or white wine respectively. Both Alice and Bob share the same interest in the wine matching the dinner, but while Alice knows what she is preparing for dinner, Bob does not. However, Bob does not need to guess what Alice is preparing because Alice can simply tell him by saying "I'm preparing beef/fish." Only then would Bob make his decision to bring either red or white wine.

This contrived scenario is perhaps the simplest possible example of a stereotypical informative assertion: the speaker (Alice) has some piece of information that the hearer (Bob) lacks but would like to have in order to make a well-informed decision; the speaker then utters a sentence (which we may assume has a semantic meaning already) and the hearer possibly changes his initial beliefs in some fashion and chooses his action subsequently. This idealized situation should then be modelled as a particular dynamic game with incomplete information.

SIGNALING GAMES. The crucial ingredients of the context of utterance of the previous example —such as Alice's knowledge of what she is preparing for dinner; Bob's uncertainty thereof; Alice's and Bob's reasonably available choices; their desires and preferences— all can be captured in a relatively simple game called a SIGNALING GAME. A signaling game is a special kind of dynamic game with incomplete information that has been studied extensively in philosophy (Lewis 1969), economics (Spence 1973), biology (Zahavi 1975; Grafen 1990) and linguistics (Parikh 1991, 1992, 2001; van Rooij 2004*b*).[12]

12. Parikh explicitly denies that he is using standard signaling game models. Though fairly similar to signaling games, his GAMES OF PARTIAL INFORMATION are not quite the same and are also not standard in game theory (see Parikh 2006).

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

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

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 full-support probability distribution over T, which usually represents the receiver's uncertainty which state in T is actual;[13] 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 set of all states where that message is true (or otherwise semantically acceptable); 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 that give a numerical value for, roughly, the desirability of each possible play of the game.[14]

The states of a signaling game basically fix which utility-relevant results the players' actions will have. When specifying utility functions for sender and receiver, it is then often convenient to distinguish what part of a player's payoff results from the sender's choice of a message and what part results from the receiver's choice of action. For instance, we sometimes like to think of utility functions $\mathrm{U}_{S,R}$ as composed of RESPONSE UTILITIES $\mathrm{V}_{S,R} : T \times A \to \mathbb{R}$ minus MESSAGE COSTS $\mathrm{C}_{S,R} : T \times M \to \mathbb{R}$, so that:

$$\mathrm{U}_{S,R}(t, m, a) = \mathrm{V}_{S,R}(t, a) - \mathrm{C}_{S,R}(t, m).$$

This makes it easier to express ideas such as that messages are entirely *costless*: $\mathrm{C}_{S,R}(t, m) = 0$ for all t and m. Whenever messages are costless we speak of CHEAP TALK.

THE WINE-CHOICE SIGNALING GAME. The above wine-choice scenario can be represented as the signaling game given in figure 1.2. There are two possible states of nature (only one of which is actual, of course): in $t_{\texttt{beef}}$ Alice prepares beef, and in $t_{\texttt{fish}}$ she prepares fish. Alice knows which state is actual, but Bob does not and so his uncertainty is represented numerically in the probability

13. 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, $X : Y \to Z$ is alternative notion for $X \in Z^Y$, and $\mathcal{P}(X)$ is the power set of X. We say that a probability distribution $\delta \in \Delta(X)$ has FULL SUPPORT if for all $x \in X$ $\delta(x) > 0$. To ask for full support receiver beliefs is to require that the receiver does not rule out *a priori* that certain states are actual, which is fairly natural.

14. I will assume implicitly throughout this thesis that signaling games also satisfy a minimal condition on expressibility, namely that for each state t there is at least one message m such that $t \in [\![m]\!]$.

		U_S, U_R		M & $[\![\cdot]\!]$	
	$\Pr(t)$	$a_{\texttt{red}}$	$a_{\texttt{white}}$	$m_{\texttt{beef}}$	$m_{\texttt{fish}}$
$t_{\texttt{beef}}$	$3/5$	1,1	0,0	$\surd$	–
$t_{\texttt{fish}}$	$2/5$	0,0	1,1	–	$\surd$

Figure 1.2: Signaling game for coordination

distribution Pr. According to the table in figure 1.2 then, Bob finds it just a little more likely that Alice prepares beef than that she prepares fish (perhaps because she has shown a tendency towards beef in the past). Alice can say either $m_{\texttt{beef}}$ "I'm preparing beef" or $m_{\texttt{fish}}$ "I'm preparing fish" with the obvious semantic meaning as indicated by the check marks in figure 1.2: a check mark in the table means that a message is true in a given state. Bob can then choose to bring red wine ($a_{\texttt{red}}$) or white wine ($a_{\texttt{white}}$). Both Alice and Bob value an outcome where the wine matches the food more than an outcome where it doesn't; beyond that, they have even identical preferences in the given example. The game is one of pure cooperation and coordination, as preferences are aligned, and as states and response actions have to be matched in order to obtain maximal payoffs.

A few remarks are in order with respect to this example. First of all, it should be mentioned that wherever utilities are given as in the table in figure 1.2, it is implicitly assumed that messages are costless. Secondly, we should also notice that the simple signaling game as defined here *does* allow the speaker to send untrue messages: although the semantic meaning of messages is represented in $[\![\cdot]\!]$, the sender is not forbidden to send, say, $m_{\texttt{beef}}$ in state $t_{\texttt{fish}}$. There are other ways of defining the signaling game model for the wine-choice scenario, of course, and indeed we will come back at length to the issue of message costs, truthful sending and semantic meaning in section 1.2.4 later in this chapter.

The Some-All Game. Of course, the signaling game in figure 1.2 is not particularly interesting for linguistic pragmatics. There is not much room for pragmatic inference in this toy example: commonsense has it that Alice would tell Bob that she is preparing beef if and only if she is indeed preparing beef, and Bob will bring red wine if and only if Alice tells him that she is preparing beef. A context model of a pragmatically more interesting situation is the signaling game in figure 1.3, which is intended to capture in abstraction

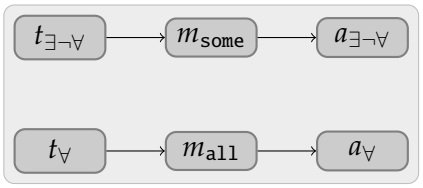
	$\Pr(t)$	$a_{\exists\neg\forall}$	$a_{\forall}$	$m_{\texttt{some}}$	$m_{\texttt{all}}$
$t_{\exists\neg\forall}$	$1-p$	1,1	0,0	$\surd$	–
$t_{\forall}$	p	0,0	1,1	$\surd$	$\surd$

Figure 1.3: The some-all game: a context model for scalar implicature

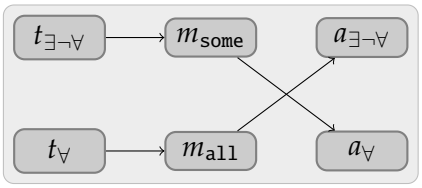
the arguably simplest context of utterance in which we would expect a scalar implicature like the one in (3) to arise.[15] The signaling game in figure 1.3 has two states $t_{\exists\neg\forall}$ and $t_{\forall}$, two messages $m_{\texttt{some}}$ and $m_{\texttt{all}}$ with semantic meaning as indicated and two receiver interpretation actions $a_{\exists\neg\forall}$ or $a_{\forall}$ which correspond one-to-one with the states. We could think of these actions either as concrete actions, as interpretations that the receiver wants to adopt or just as placeholders indicating what is relevant for the receiver in the given context. Also in this example sender and receiver payoffs are perfectly aligned in order to model the assumption that interlocutors cooperate and care to coordinate on proper interpretation.

Pure Strategies Capture Agent Behavior. Recall that games specify the general behavioral possibilities of agents, but do not specify further how agents do or should in fact behave. Signaling games therefore are also merely models of the context of utterance, but not of the behavior of agents. In general, behavior of players in dynamic games is represented in terms of strategies which select possible moves for each agent for any of their choice points in the game. For signaling games, a pure sender strategy $s \in M^T$ is a function from states to messages which specifies which message the sender will or would send in each state that might become actual. A pure receiver strategy $r \in A^M$ is a function from messages to actions which similarly specifies which action the receiver will or would choose as a response to each message he might observe. (Obviously, the receiver knows only what message has been sent, but not what state is actual, so he has to choose an action for each message he might observe and cannot condition his choice on the actual state

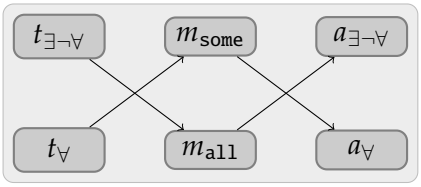
15. The reader is asked to bear with my choice of context models until I have had a chance to motivate my basic assumptions about signaling games for pragmatic interpretation in chapter 3. For the time being, suffice it to say that a signaling game like that in figure 1.3 should be thought of as the hearer's construction of a default context for the interpretation of a sentence like (8a) that he constructs *after* hearing the target utterance. This motivates, for example, omission of a state $t_{\neg\exists}$ and yields an interpretation of prior probabilities that legitimates the assumption of (mostly) flat priors.



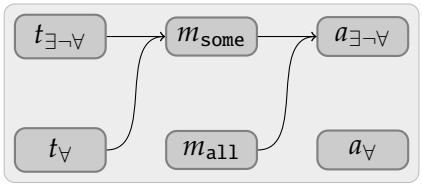
(a) Scalar implicature



(b) Unreasonable behavior



(c) Permuted meaning



(d) Pooling

Figure 1.4: Some pure strategy profiles for the some-all game

of affairs). A PURE STRATEGY PROFILE $\langle s, r \rangle$ is then a characterization of the players' *joint behavior* in a given signaling game, and the set of all such pairs gives the set of all behavioral possibilities of our abstract conversationalists.

A strategy profile can be represented as in figure 1.4, where four out of the sixteen possible pure strategy profiles of the some-all game are given. Sender strategies (functions in M^T) are represented by the set of arrows leaving the state nodes on the left; receiver strategies (functions in A^M) are represented by the set of arrows leaving the message nodes in the middle. Under the convention that the nodes representing states, messages and actions are arranged as in figure 1.4, we can represent the set of all sixteen possible pure strategy profiles of the some-all game perspicuously as in figure 1.5. For clarity, the strategy profiles in figures 1.4a, 1.4b, 1.4c and 1.4d have numbers 1, 13, 16, and 6 in figure 1.5 respectively.

SOLUTIONS AND PRAGMATIC EXPLANATIONS. In a situation modelled by the some-all game in figure 1.3, we would intuitively expect the sender and receiver to behave as described by the strategy profile in figure 1.4a: (i) the sender sends $m_{\texttt{some}}$ in state $t_{\exists\neg\forall}$ and the message $m_{\texttt{all}}$ in state $t_{\forall}$; and (ii) the receiver responds to $m_{\texttt{some}}$ with $a_{\texttt{some}}$ and to $m_{\texttt{all}}$ with $a_{\texttt{all}}$. This would correspond to the intuitive use of the corresponding natural language expressions. If for a given solution concept for signaling games the intuitive strategy profile in figure 1.4a was an accepted solution, and if no other strategy profile

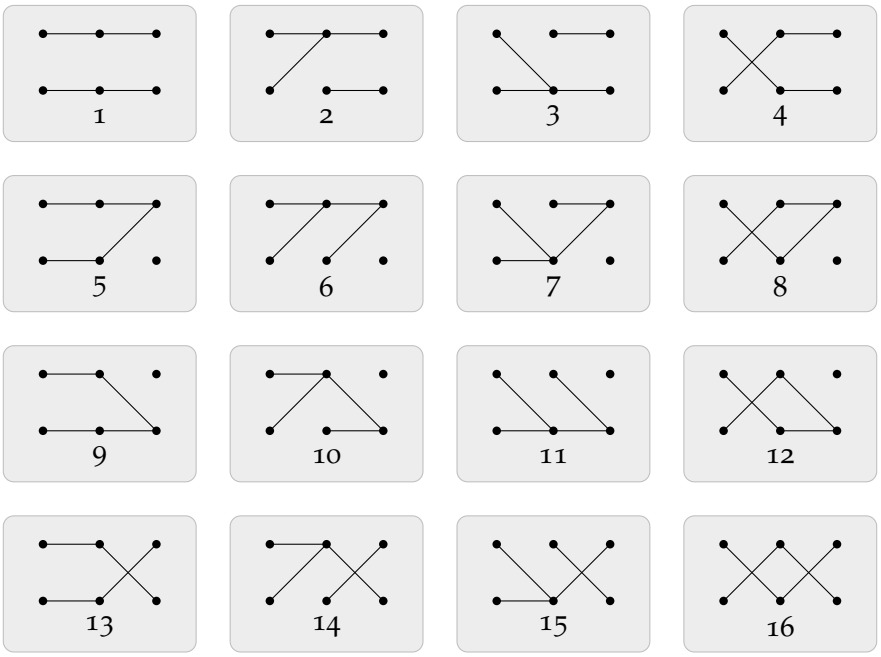


Figure 1.5: All pure strategy profiles for the some-all game

was, then this would, in a sense, *explain* the scalar implicature.

For this kind of explanation of the pragmatic data, game theoretic pragmatics generally requires us to do two things: firstly, we need to set up a signaling game as a reasonable and sufficient representation of the context of utterance of a sentence whose use and pragmatic interpretation we would like to explain; secondly, an appropriate solution concept should then select all and only those strategy profiles that represent the intuitively or empirically attested data. Together, we would then regard the pair consisting of the game-as-context-model and the solution concept as the explanation of the data.

As for the game model, I will be using signaling games as a model of utterance context throughout this thesis (whence also the title). I made the case that a conversational move ought to be modelled as *some* dynamic game of incomplete information. My focus on signaling games is for mere conceptual convenience: these models are just the simplest non-trivial models in which we can study pragmatic phenomena. As for the solution concept, there are at least two requirements to be met. First and foremost, we would like a solution concept that given a reasonable representation of the context *uniquely* selects the adequate strategy profile. But on top of that, since dif-

ferent solution concepts can also have quite different conceptual justifications and epistemic interpretations —as we have already seen above in the context of static games— we do not just want *any* solution concept that yields the right behavioral predictions; we would prefer a notion that pays due respect to agents' reasoning about the context and each others' (belief in) rationality in a psychologically plausible and preferably empirically vindicated way.

This is basically what the first part of this thesis is trying to achieve. In order to meet this challenge I will take an explicitly epistemic approach to game theory. The solution concept that I will offer in chapter 2 will crucially rely on assumptions about the cognitive architecture —including reasonable limitations— of language-using agents. To pave the way, the following section spells out the necessary basic notions of behavior, rationality and beliefs in signaling games that underlie the definitions and interpretations of different solution concepts.

1.2.3 Solving Signaling Games

Strategies

We said that a *pure sender strategy* in a signaling game is a function $s \in M^T$ and that a *pure receiver strategy* is a function $r \in A^M$. Let S and R be the sets of all pure sender and receiver strategies. Pure strategies define how a player behaves in each possible information state that she might find herself in during the game.

Probabilistic Strategies. Next to pure strategies, there are also two kinds of probabilistic strategies: (i) mixed strategies and (ii) behavioral strategies. These should, strictly speaking, be distinguished formally and conceptually although they are equivalent in the context of signaling games. A mixed strategy is a probability distribution over the set of pure strategies. So a mixed sender strategy is a probability distribution $\tilde{s} \in \Delta(M^T)$; and a mixed receiver strategy is a probability distribution $\tilde{r} \in \Delta(A^M)$. A behavioral strategy is a map from information states of a player to a probability distribution over possible moves in that information state. So a behavioral sender strategy is a function $\sigma \in \mathcal{S} = (\Delta(M))^T$ and a behavioral receiver strategy is a function $\rho \in \mathcal{R} = (\Delta(A))^M$.

However, for games with perfect recall, in which players never forget any information that they had at previous information states (see Osborne and Rubinstein 1994, pp. 203–204), the two ways of specifying probabilistic

strategies are equivalent in the sense that they give rise to the same probability distribution over outcomes (see Osborne and Rubinstein 1994, pp. 212–216). This result is known as 'Kuhn's theorem.' Evidently, signaling games satisfy the perfect-recall requirement trivially, because each player only moves once and can thus never forget previously held information.[16]

Interpretation of Probabilistic Strategies. The proper interpretation of probabilistic strategies has been a matter of engaged debate among game theorists (see Osborne and Rubinstein 1994, pp. 37–44). Does the assumption of probabilistic strategies mean that we want players to intentionally randomize their actions (in certain, possibly restricted ways) instead of selecting a concrete single move? To many this seems a dubious design in many contexts, and it certainly seems strange in the context of natural language use and interpretation (but see also Aumann 1974). Another widely held and reasonable interpretation of especially mixed strategies is as the frequentist probability of pure strategies occurring in a population of players. This is a particularly appealing interpretation for equilibrium-based concepts in a diachronic, evolutionary approach. Still, in the present context my preferred interpretation of probabilistic strategies is as *conjectures* about the behavior of the opponent: rather than allowing players to randomize at will, choices remain finite and concrete; but other players may not know what their opponent is playing, so that their uncertainty about the concrete strategy played by an opponent is represented by a probabilistic strategy.

If interpreted as a belief about an opponent's move, the representation of probabilistic strategies as behavioral strategies, rather than mixed strategies may seem more intuitive: instead of having a conjecture about the completely specified *conditional behavior* of an opponent, agents have *conditional conjectures* about what the opponent will do in each situation where she is called to act. That is why, in the context of signaling games, I will stick to behavioral strate-

16. The following straightforward conversions yield the desired outcome equivalence. Given a mixed sender strategy $\tilde{s} \in \Delta(M^T)$, an equivalent behavioral sender strategy σ is:

$$\sigma(m|t) = \sum_{\{s \in \mathsf{S} \mid s(t)=m\}} \tilde{s}(s).$$

Given a behavioral sender strategy σ an equivalent mixed strategy is $\tilde{s}$ is given by:

$$\tilde{s}(s) = \prod_{t \in T} \sigma(s(t)|t).$$

The conversion of probabilistic receiver strategies is analogous.

gies as a representation of probabilistic strategies which, in turn, represent uncertainty about the opponent's behavior.

Beliefs

Sender Beliefs. When it is her turn to act in a signaling game, the sender knows the actual state, but she does not know what the receiver will do: in technical terms, the sender has complete information (about the game situation), but imperfect information (about her opponent's behavior). Sender beliefs are then given as the set of probabilistic receiver strategies:

$$\Pi_S = \mathcal{R} = (\Delta(A))^M.$$

A given sender belief $\rho \in \Pi_S$ specifies a probability distribution over A for each m: $\rho(m)$ then gives the probabilistic beliefs of the sender about which action the receiver will play if he observes m. This is the only game-relevant uncertainty of the sender that we need to represent.

Receiver Beliefs. The situation of the receiver is a little more complicated, because the receiver not only has imperfect information (not knowing what the sender does), but also incomplete information (not knowing what the actual state of the world is). With some redundancy, we could say that there are three things that the receiver is uncertain about:

(i) R has *prior uncertainty* about which state is actual before he observes a message; these prior beliefs are specified by the distribution Pr in the signaling game;

(ii) R also is uncertain about the sender's behavior; we can thus characterize the receiver's behavioral beliefs about sender behavior as a probabilistic sender strategy, i.e., a function: $\sigma \in (\Delta(M))^T$ that gives a probability distribution over M for each t;

(iii) and finally R also has *posterior uncertainty* about which state is actual after he observes a message; for clarity, this is not because the actual state changes, but because the receiver's beliefs about the actual state may be influenced by the observation of which message the sender has sent; these posterior beliefs can be described as a function $\mu \in (\Delta(T))^M$ that gives a probability distribution over T for each m.

Taken together, the set of relevant receiver beliefs Π_R is the set of all triples $\langle \Pr, \sigma, \mu \rangle$ for which $\sigma \in (\Delta(M))^T$ and $\mu \in (\Delta(T))^M$.

Consistency. This characterization of the receiver's uncertainty is partially redundant, because there is a strong intuitive sense in which the posterior beliefs μ should be derived, at least in part, from the other two components of R's uncertainty. What we want is a further *consistency requirement* that the receiver's posterior beliefs fit his prior beliefs and his conjecture about the sender's behavior. Technically speaking, we want the posterior beliefs μ to be derived from Pr and σ by Bayesian conditionalization. We say that the receiver's posterior beliefs μ are consistent with his beliefs Pr and σ if and only if for all t in T and for all m in the image of σ, i.e., all m for which $\sigma(m|t) \neq 0$ for some t, we have:

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

Consistency effectively demands reasonable, i.e., 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. We will come back to consistency later in section 2.2.3 in the context of an example that shows it at work. For the time being, suffice it to say that consistency of beliefs is the normative standard for agent's belief formation adopted in game theory.

Surprise Messages and Counterfactual Beliefs. It's crucial to keep in mind that consistency only applies to messages in the image of σ, i.e., to messages that are expected to be sent under the belief σ. It could happen that the receiver does not expect a certain message m to be sent, in which case he would hold a belief $\langle \Pr, \sigma, \mu \rangle$ for which $\sigma(m|t) = 0$ for all states $t \in T$. Given such a belief σ the message m is a surprise message, in the sense that the receiver is (or would be) surprised if he were to observe it, as he did not expect it to be sent. Consistency, however, is a condition on non-surprise messages only; it does not restrict the receiver's counterfactual beliefs, as we could call them, defined as those beliefs he holds after surprise messages.[17]

Rationality

Bayesian Rationality. The notion of rationality in both classical decision and game theory where agents have to make decisions under uncertainty

17. I call these beliefs counterfactual because they are of the form: "S does not send m, but if she *would*, the actual state *would* be t with probability p."

about the outcomes of their actions is BAYESIAN RATIONALITY. The idea behind Bayesian rationality is maximization of EXPECTED UTILITY, which is a technical measure for the gain an action is subjectively expected to yield. Towards a general definition, fix a set of alternative actions A, and a set of states T that the outcome of performing an action depends on. We assume that our decision maker has preferences over all outcomes, i.e., pairs $T \times A$, which is given by the numerical utility function $\mathrm{U} : T \times A \to \mathbb{R}$. We also assume that she has beliefs about the actual state, which is given by a probability distribution over states $\Pr \in \Delta(T)$. The agent's *expected utility* of performing an action a as a function of belief Pr is then defined as

$$\mathrm{EU}(a, \Pr) = \sum_{t \in T} \Pr(t) \times \mathrm{U}(t, a).$$

This helps define Bayesian rationality as follows:[18]

(13) BAYESIAN RATIONALITY Given an agent's behavioral alternatives A, his beliefs Pr and preferences U, the agent is *rational* only if he chooses an action $a \in A$ which maximizes his expected utility (as given by Pr and U).

RATIONAL BEHAVIOR & BEST RESPONSES. Since a signaling game gives us the players' action alternatives (sets M and A) and the agents' preferences over outcomes (functions U_S and U_R), all we need to add to define rational behavior is a specification of the agents' beliefs in the game.

The sender's beliefs $\rho \in \Pi_S$ are probabilistic receiver strategies. Given such as belief ρ about the receiver's behavior, we can define the sender's *expected utility* of sending message m in state t as a function of her belief ρ as follows:

$$\mathrm{EU}_S(m, t, \rho) = \sum_{a \in A} \rho(m, a) \times \mathrm{U}_S(t, m, a).$$

In line with Bayesian rationality, if S is rational and believes ρ she should send a message m in state t only if it maximizes her expected utility given belief ρ. We say that a pure sender strategy $s \in \mathsf{S}$ is rational just in case it selects

18. The definition in (13) has only "only if", because, strictly speaking, an agent who chooses an act that maximizes expected utility need not *be* rational, although, in a sense, she would certainly *behave* rationally. With locution "behaves rationally" instead of "be rational" in (13) both directions of implication are true. However, since we always only reason *from* the assumption of an agent's *de facto* rationality, and not *to* it, we only need that rationality implies utility maximization in expectation.

an action which maximizes expected utility in all states, i.e., s is a RATIONAL PURE SENDER STRATEGY given belief ρ if and only if for all t:

$$s(t) \in \arg\max_{m \in M} \mathrm{EU}_S(m, t, \rho).$$

Synonymously, we say that s is a (PURE) BEST RESPONSE to belief ρ. The set of all such pure best responses to belief ρ is denoted by $\mathrm{BR}(\rho)$.

The receiver's beliefs are triples $\langle \Pr, \sigma, \mu \rangle \in \Pi_R$, but the important component for a characterization of rational receiver behavior is, of course, the posterior beliefs μ: it's *after* observing a message that the receiver is called to act, so it's with respect to the beliefs he holds at that time that we should judge him rational or not. Therefore, given a posterior belief μ, we define R's expected utility of performing a after message m has been received as

$$\mathrm{EU}_R(a, m, \mu) = \sum_{t \in T} \mu(t|m) \times \mathrm{U}_R(t, m, a)$$

and say that $r \in \mathsf{R}$ is a RATIONAL PURE RECEIVER STRATEGY if and only if for all m

$$r(m) \in \arg\max_{a \in A} \mathrm{EU}_R(a, m, \mu).$$

Alternatively, we call such an r a (PURE) BEST RESPONSE to belief π_R (or, simply, to μ). The set of all such pure best responses to belief π_R is denoted by $\mathrm{BR}(\pi_R)$ (or, sometimes, $\mathrm{BR}(\mu)$).

Equilibrium & Rationalizability

Having defined what behavior is rational for sender and receiver individually, we are able to define basic solution concepts for signaling games, in particular equilibrium and rationalizability.

RATIONALIZABILITY. Rationalizability aims to single out behavior that is (i) rational and (ii) consistent with a belief in common belief in rationality. Remember from strategic games that the algorithmic idea behind rationalizability is that of iteratively eliminating strictly dominated strategies. For signaling games this would mean that starting from the set of all pure sender and receiver strategies, we would like to rule out iteratively all those pure strategies which are strictly dominated, i.e., which are never a best response to any belief in the remaining opponent strategies.

Towards a formal definition, recall that S and R are the sets of pure sender and receiver strategies. Let us fix $\mathsf{S}_0 = \mathsf{S}$ and $\mathsf{R}_0 = \mathsf{R}$, and then define

inductively the sets S_{n+1} and R_{n+1} of pure sender and receiver strategies in S_n and R_n respectively that are rational given some belief in R_n and S_n, i.e., some belief that the opponent plays some strategy in the set R_n or S_n. Formally, the induction step reads as:

$$\begin{aligned}
\mathsf{S}_{n+1} &= \{s \in \mathsf{S}_n \mid \exists\, \rho \in \Delta(\mathsf{R}_n) \;:\; s \in \mathrm{BR}(\rho)\} \\
\mathsf{R}_{n+1} &= \{r \in \mathsf{R}_n \mid \exists\, \pi_R = \langle \Pr, \sigma, \mu\rangle \in \Pi_R \;: \\
&\qquad\qquad \text{(i)} \quad r \in \mathrm{BR}(\mu) \\
&\qquad\qquad \text{(ii)} \quad \pi_R \text{ is consistent} \\
&\qquad\qquad \text{(iii)} \quad \sigma \in \Delta(\mathsf{S}_n)\ \}.
\end{aligned}$$

Finally, the sets of RATIONALIZABLE STRATEGIES are the sets

$$\mathrm{Rat}_S = \bigcap_{i \in \mathbb{N}} \mathsf{S}_i \qquad\qquad \mathrm{Rat}_R = \bigcap_{i \in \mathbb{N}} \mathsf{R}_i.$$

The set Rat_S is the set of all pure sender strategies which are compatible with the assumption that S is rational and believes in common belief in rationality. The same holds for the receiver, of course. For a strategy profile $\langle s, r\rangle$ to be rationalizable it suffices for s and r to be rationalizable. That means that rationalizability does not require beliefs about opponent strategies to be correct: rationalizability is a non-equilibrium solution concept.

Without any further restrictions, rationalizability is a fairly weak solution concept.[19] For instance, if we assume that talk in the some-all game in figure 1.3 is cheap and that $\Pr(t_{\exists\neg\forall}) = \Pr(t_{\forall})$, then *any* pure strategy profile is rationalizable, because any possible sender or receiver strategy can be rationalized by a belief in some opponent behavior. As a solution concept for (cheap talk) signaling games in game theoretic pragmatics this basic version of rationalizability therefore is far too unrestricted. This is a negative, but as such noteworthy result: in cheap talk signaling games like the some-all game the assumption that agents are rational and believe in common belief in rationality is *not* enough to explain pragmatic language use and interpretation.

PERFECT BAYESIAN EQUILIBRIUM. As we have seen for strategic games above, an equilibrium solution concept characterizes a mutually optimal, hence steady, pattern in the joint behavior of players. A set of strategies is in equilibrium if nobody has an incentive to deviate given that everybody else conforms. Thus, equilibrium requires that the beliefs of players be correct, i.e., derived from

19. We will come back to stronger notions of rationalizability in section 2.4.3.

the strategy profile (at least as far as possible) and that each individual is responding rationally to that belief. Equilibrium does *not* require belief in the opponent's rationality.

For signaling games, this comes down to saying that the pure strategy profile $\langle s, r \rangle$ is in equilibrium just in case (i) s is rational given the belief that the receiver plays r and (ii) r is rational given the belief that the sender plays s. More precisely, the proper general definition is in terms of probabilistic strategies. We say that a triple $\langle \sigma, \rho, \mu \rangle$ is a PERFECT BAYESIAN EQUILIBRIUM (PBE) iff three conditions hold:[20]

(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.

In order to check whether a given pure strategy profile $\langle s, r \rangle$ is a PBE, we then need to consult the triple $\langle \sigma, \rho, \mu \rangle$ where σ and ρ are the unique probabilistic strategies corresponding to the pure strategies s and r, and where μ is some appropriate posterior belief of the receiver.

EXAMPLE: EQUILIBRIA OF THE SOME-ALL GAME. To illustrate the concept of perfect Bayesian equilibrium, let us briefly turn to the question which strategy profiles of the some-all game in figure 1.3 are PBEs. For the time being, let us assume that (i) we have flat prior probabilities, i.e., $\Pr(t_{\exists\neg\forall}) = \Pr(t_{\forall})$ and that (ii) the utilities given are response utilities, so that talk is cheap, i.e., that all messages can be used at no cost in all states whether they are true or not. Under these conditions all the strategy profiles highlighted in figure 1.6 are PBEs. I will not give arguments for all of the sixteen strategy profiles, but focus for the purpose of illustration on the four strategy profiles given in figure 1.4, i.e., numbers 1, 13, 16, and 6 in figure 1.6.

To begin with, let us check that strategy profile number 1, which is the intuitive play in figure 1.4a, is a PBE. Let then σ_1 and ρ_1 be the relevant

20. Strictly, speaking, we have so far only defined rationality for *pure* strategies. Say that a *probabilistic* strategy X, be it sender's or receiver's, is rational given belief π iff, when considered a mixed strategy, X puts positive probability only on pure best responses to π.

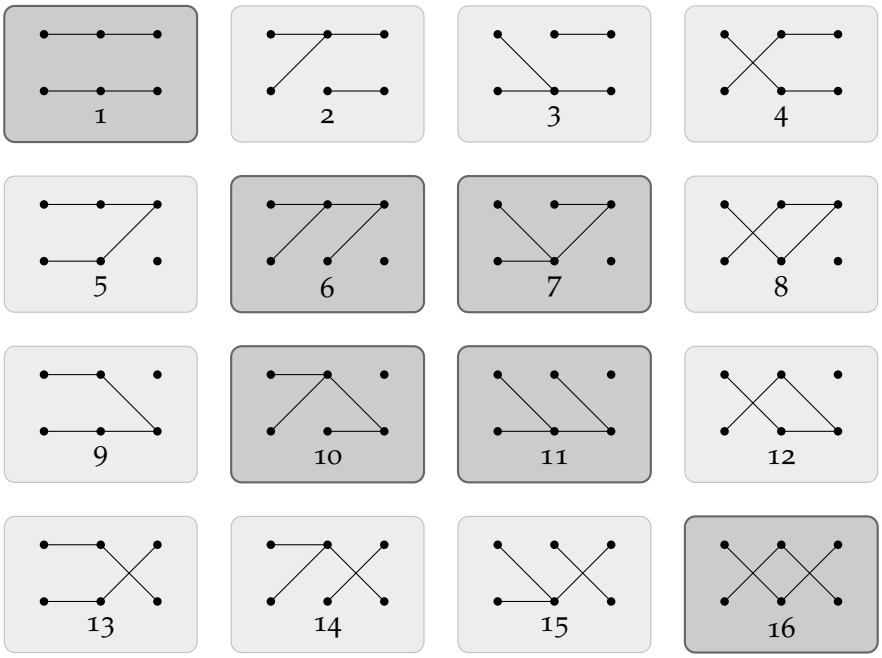


Figure 1.6: Perfect Bayesian equilibria of the some-all game (assuming cheap talk and flat priors)

probabilistic strategies.

$$\sigma_1 = \begin{bmatrix} t_{\exists\neg\forall} & \mapsto & \begin{bmatrix} m_{\texttt{some}} & \mapsto & 1 \\ m_{\texttt{all}} & \mapsto & 0 \end{bmatrix} \\ t_{\forall} & \mapsto & \begin{bmatrix} m_{\texttt{some}} & \mapsto & 0 \\ m_{\texttt{all}} & \mapsto & 1 \end{bmatrix} \end{bmatrix}$$

$$\rho_1 = \begin{bmatrix} m_{\texttt{some}} & \mapsto & \begin{bmatrix} a_{\exists\neg\forall} & \mapsto & 1 \\ a_{\forall} & \mapsto & 0 \end{bmatrix} \\ m_{\texttt{all}} & \mapsto & \begin{bmatrix} a_{\exists\neg\forall} & \mapsto & 0 \\ a_{\forall} & \mapsto & 1 \end{bmatrix} \end{bmatrix}$$

It is obvious that the only sender strategy which is rational given S's preferences U_S and the belief ρ_1 is σ_1. Moreover, the receiver's posterior beliefs are completely determined by the sender's strategy σ_1: the only belief μ_1 consistent with any full support prior and the behavioral belief σ_1 is the posterior belief that puts full credence, i.e., probability 1, on state $t_{\exists\neg\forall}$ after hearing

$m_{\texttt{some}}$ and full credence on $t_\forall$ after hearing $m_{\texttt{all}}$:

$$\mu_1 = \begin{bmatrix} m_{\texttt{some}} & \mapsto & \begin{bmatrix} t_{\exists\neg\forall} & \mapsto & 1 \\ t_\forall & \mapsto & 0 \end{bmatrix} \\ \\ m_{\texttt{all}} & \mapsto & \begin{bmatrix} t_{\exists\neg\forall} & \mapsto & 0 \\ t_\forall & \mapsto & 1 \end{bmatrix} \end{bmatrix}.$$

But then, given μ_1 and the receiver's preferences U_R, ρ_1 is indeed rational, in fact the only rational receiver strategy. Consequently, the tuple $\langle \sigma_1, \rho_1, \mu_1 \rangle$ is a PBE. This is as it should be, for a game theoretic explanation of the scalar implicature.

Similarly, it turns out that the strategy profile number 13, given in figure 1.4b, is *not* a PBE: informally speaking, if the sender's strategy σ_{13} reveals the actual state, it is irrational for the receiver to reverse the meaning of the signals. This, too, is a welcome prediction of perfect Bayesian equilibrium, for intuitively the strategy profile number 13 *should* be ruled out.

However, unfortunately, the strategy profiles numbers 16 and 6, given also in figures 1.4c and 1.4d, which also represent intuitively unattested kinds of conversational behavior, are *not* ruled out by our solution concept as it stands: profiles 1.4c and 1.4d are PBEs in the cheap talk some-all game. The interested reader will quickly verify for herself that number 16 is. The argument why the strategy profile number 6 is a PBE too is slightly more complicated and it pays to briefly enlarge on it here.

Let σ_6 and ρ_6 be the relevant probabilistic strategies. It is important to notice that, unlike in all other cases so far, there is not just one receiver belief consistent with the behavioral belief σ_6. Indeed, *any* belief $\langle \Pr, \sigma_6, \mu_6^q \rangle$ with posterior belief

$$\mu_6^q = \begin{bmatrix} m_{\texttt{some}} & \mapsto & \begin{bmatrix} t_{\exists\neg\forall} & \mapsto & \Pr(t_{\exists\neg\forall}) \\ t_\forall & \mapsto & \Pr(t_\forall) \end{bmatrix} \\ \\ m_{\texttt{all}} & \mapsto & \begin{bmatrix} t_{\exists\neg\forall} & \mapsto & q \\ t_\forall & \mapsto & 1-q \end{bmatrix} \end{bmatrix}$$

with $q \in [0,1]$ is consistent. However, not every such posterior belief μ_6^q makes ρ_6 a rational receiver strategy. First of all, ρ_6 can only ever be rational if $\Pr(t_{\exists\neg\forall}) \geq 1/2$. In other words, only for some versions of the some-all game can the profile number 6 be a PBE. Moreover, the receiver strategy ρ_6

is rational only for values $q \geq 1/2$. That means that *not all* consistent beliefs make the given pure strategy profile a PBE. Nonetheless, there are posteriors which fulfill the requirements of perfect Bayesian equilibrium together with σ_6 and ρ_6, so that we count the strategy profile number 6 as among the PBEs. (We will come back to this kind of slack in the receiver's counterfactual beliefs at various points throughout the thesis.)

Taken together, if we assume that talk is cheap, we find that some of the unintuitive strategy profiles are ruled out by perfect Bayesian equilibrium, but not all of them. This is obviously not a satisfactory result for game theoretic pragmatics where we would like to single out the strategy profile number 1 uniquely. As I have argued before, this problem of finding a proper solution concept will indeed be our foremost challenge in game theoretic pragmatics. But there are others too, some of which this thesis will try to meet. The next section addresses these challenges and issues that arise for game theoretic pragmatics.

1.2.4 Implementing Semantic Meaning

Both of our basic solution concepts for signaling games, rationalizability and perfect Bayesian equilibrium, are too weak to explain pragmatic language use and interpretation in cheap talk signaling games. Conceptually this means that it is not enough to explain pragmatic behavior to just assume either, as rationalizability does, that agents behave rationally given a belief in common belief in rationality, or, as perfect Bayesian equilibrium does, that agents behave rationally given a true belief about opponent behavior. This much is indeed a conceptually interesting result: our basic notions of rational interaction alone are not enough to explain pragmatic phenomena; something else needs to be added.

In essence, this problem could be conceived of as a concrete instance of the more general PROBLEM OF EQUILIBRIUM SELECTION, well-known and notorious in game theory. In theoretical economics there is a whole branch of literature, the so-called *refinement literature*, dedicated to the search for appropriate refinements of standard equilibrium concepts, such as perfect Bayesian equilibrium. It may therefore appear fair to say that the most confronting problem of game theoretic pragmatics is, in a sense, a game theoretic one.

Nonetheless, it is clearly not very surprising that rationalizability and perfect Bayesian equilibrium yield too weak predictions for cheap talk signaling games. Obviously, what should be added is that which has so far frivolously

been left out of the picture: the conventional semantic meaning of messages. The following therefore will dwell on this issue, argue that semantic meaning should be integrated into the solution concept (as opposed to the game model), and gesture at the conceptual difficulties in doing so.

Impossible Falsity

When looking at the set of PBEs of the some-all game in figure 1.6 it strikes us that in some PBEs the sender uses messages that are false. This suggests trying to single out the intuitive profile number 1 uniquely by assuming that the sender *has to* send true messages. This would boil down to a change in the structure of the game, restricting the allowed moves of the sender. Indeed, this is what most previous work in game theoretic pragmatics assumes (e.g. Parikh 1992, 2001; Benz 2006; Benz and van Rooij 2007; van Rooij 2008).

For the some-all game this immediately rules out all those strategy profiles where message $m_{\texttt{all}}$ is sent in state $t_{\exists\neg\forall}$ (numbers 7, 11, and 16) and leaves us with the restricted set of PBEs in figure 1.7. Plainly, this pruning of the strategy space circles in on the desired solution but still is too inclusive, as the two pooling strategies numbered 6 and 10 are still PBEs. But let us briefly ask whether we cannot use the idea of impossible false signaling to restrict the set of equilibria even further.

Indeed, there is something fishy about at least the strategy profile number 6. We have seen above that receiver strategy ρ_6, which is part of strategy profile 6, is rational only for a posterior belief μ_6 for which $\mu(t_{\exists\neg\forall}|m_{\texttt{all}}) \geq 1/2$. But even if $m_{\texttt{all}}$ is a surprise message, this posterior belief should actually also be ruled out if it is part of the game structure (and hence common knowledge between players) that the sender *cannot* send untrue messages. To wit, if the sender cannot possibly send semantically untrue messages and the receiver knows this, then this knowledge should also be contained in any counterfactual beliefs of the receiver. In particular, the receiver should not believe that it is possible *at all* that the actual state is $t_{\exists\neg\forall}$ after the message $m_{\texttt{all}}$, no matter whether the receiver expects $m_{\texttt{all}}$ to be sent or not; hence, any posterior belief that faithfully reflects knowledge of the game structure would set $\mu(t_{\exists\neg\forall}|m_{\texttt{all}}) = 0$. Perfect Bayesian equilibrium, it turns out, does not restrict the receiver's counterfactual beliefs appropriately to reflect (knowledge of) the game structure.

Of course, various refinements of equilibrium exist that do take the relevant game structure sufficiently into account. Such refinements differ in

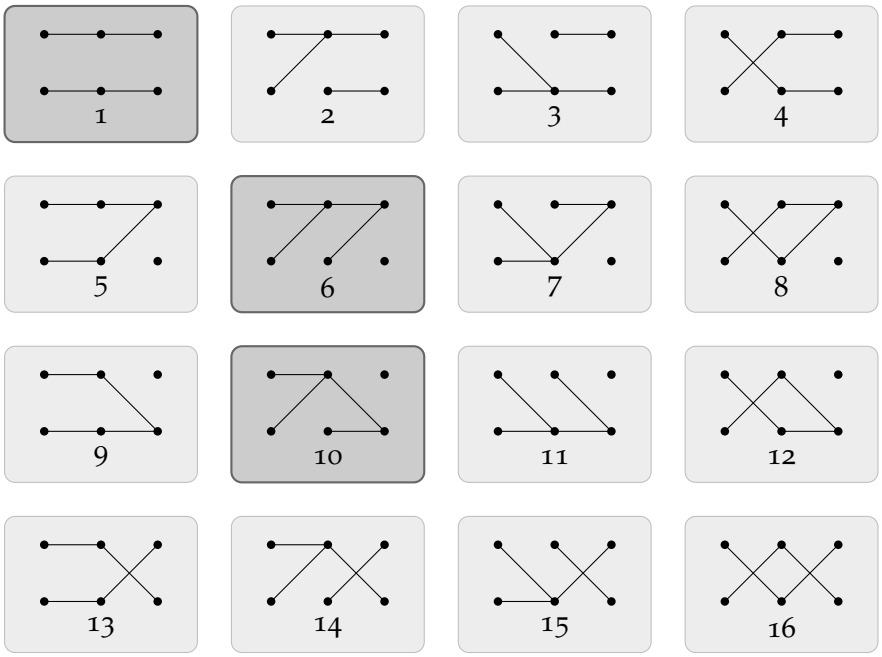


Figure 1.7: Perfect Bayesian equilibria of the some-all game (assuming flat priors and that truthful sending is obligatory)

conception and formal implementation and often tend to be mathematically quite complicated for various reasons. Let me just give two prominent examples, without going into any formal detail. One possibility is to assume that (the receiver believes that) the sender might make small mistakes in his execution of a strategy, while still being confined by the requirement to send truthfully. This is the essential idea behind TREMBLING HAND PERFECT EQUILIBRIUM (Selten 1965, 1975). Another option is SEQUENTIAL EQUILIBRIUM (Kreps and Wilson 1982) which requires, rather technically, that the receiver's posterior beliefs μ be derived from the limit of an infinite sequence of beliefs in non-pooling sender strategies. Details are inessential here (see Osborne and Rubinstein 1994, for discussion and comparison), just suffice it to say that both perfect equilibrium as well as sequential equilibrium secure that the structure of the game is taken into account in the formation of counterfactual beliefs. That means that both notions exclude strategy profile number 6 if we fix that the sender *has to* send truthfully.

Still, no matter how we might try, there is no way that a counterfactual belief in truthful sending could rule out the strategy profile number 10 as an equilibrium. So, it seems that making truth obligatory in the available sender choices does not quite solve the problem of equilibrium selection even if we

turn to further refinement notions.

A further conceptual problem is that although it might make sense at first glance to assume truthful signaling in cases of pure coordination, it is not reasonable to assume *in the context model* that the speaker *cannot* —not even for fun, so to speak— use a signal that is not true. I can very well say whatever I like, whenever I like to whomever I like. I may have to face social or even legal consequences from time to time, but it is not as if the semantics of my language restricts the muscles of my jaw and vocal tract, regulating what I possibly can and what I cannot utter.

Penalized Falsity

The idea that it is considerations of social or legal appropriateness that regulate what to say and what not to say suggests that we might want to implement the semantic meaning of messages as a norm, infringement of which (probably or actually) incurs a cost for the speaker. According to this approach, rather than plainly impossible, it would sometimes be irrational to send untrue signals. Conceptually speaking, this seems a more realistic design choice than to rule out false signaling altogether.

To see what implications costs for untruthful signaling have, let us return to the some-all game in figure 1.3 and drop the assumption that talk is cheap. Let us assume that the utilities given there are response utilities and that the sender's overall utilities are computed by subtracting from her response utilities a fixed penalty $c > 0$ whenever she sends a message m in a state where m is not true.

How big should the penalty c be? First of all, in order to rule out the unintuitive profile 16, we need to chose $c > 1$. This is readily verified by acknowledging that if $c < 1$ the sender who believes in ρ_{16} would rather incur her cost in order to coordinate on proper interpretation of the (false) message $m_{\texttt{all}}$ in state $t_{\exists\neg\forall}$; if $c = 1$ the sender is indifferent, and so it is still rational to use $m_{\texttt{all}}$ in state $t_{\exists\neg\forall}$. But then, even for $c > 1$ we cannot rule out strategy profiles 6 and 10 either: we are in the exact same situation as with strictly impossible false signaling. Indeed, the parallelism continues, since we could in principle rule out strategy profile 6 with a suitable refinement that restricts the receiver's counterfactual beliefs in such a way as to reason that it *would be* irrational to send $m_{\texttt{all}}$ in state $m_{\exists\neg\forall}$ even when $m_{\texttt{all}}$ is not expected in the first place.[21] But again no such refinement that includes proper rationality

21. This is indeed what the *intuitive criterion* of Cho and Kreps (1987) would give us. I will

	$a_{\texttt{heads}}$	$a_{\texttt{tails}}$	$m_{\texttt{heads}}$	$m_{\texttt{tails}}$
$t_{\texttt{heads}}$	0, 1	1, 0	$\surd$	$-$
$t_{\texttt{tails}}$	1, 0	0, 1	$-$	$\surd$

Figure 1.8: Matching pennies signaling game

considerations into the receiver's counterfactual belief formation would be able to rule out strategy profile 10.

It transpires that penalizing untrue signaling is no more useful than making it entirely impossible: though perhaps conceptually more plausible, it yields pretty much the same predictions under equilibrium notions. The problem with both impossible and penalized falsity is that these restrictions on the game model still necessitate refinements of solution concepts. But if neither of these options as such allows standard solution concepts to be used, we might as well forget about the restrictions on the game model and look for an appropriate 'semantic solution concept' in the first place. The following section gives a further argument why we should do so.

Credibility Intuitions

Consider a simple arranged situation in which Alice and Bob are playing the following game. A judge flips a fair coin and only Alice observes the outcome of the coin flip, while Bob does not. Bob has to guess the outcome of the coin flip and wins iff Alice loses iff Bob guesses correctly. But suppose that before Bob makes his guess, Alice has the chance to say "I have observed tails/head," and that it really does not matter at all whether what she says is true of false.[22] This is, in effect, a 'matching pennies'-style, zero-sum signaling game with cheap talk of the form given in figure 1.8.[23] How should Alice's announcement affect Bob's decision? It seems it shouldn't at all. Bob knows that Alice does not want to reveal anything, so neither statement should have much impact on him: we feel that Bob is well advised to just ignore what Alice says.

---

not enlarge on this here, as we will come back to such *forward induction reasoning* in section 2.3.

22. We could have Alice say whatever she wants as long as it excludes threats, bribes or promises that might alter Bob's preferences. For simplicity, we only look at these two messages.

23. We can omit listing prior probabilities when these are flat.

	$a_{\texttt{heads}}$	$a_{\texttt{tails}}$	$a_{\texttt{coop}}$	$m_{\texttt{heads}}$	$m_{\texttt{tails}}$	$m_{\texttt{coop}}$
$t_{\texttt{heads}}$	0, 1	1, 0	0, 0	$\surd$	–	–
$t_{\texttt{tails}}$	1, 0	0, 1	0, 0	–	$\surd$	–
$t_{\texttt{coop}}$	0, 0	0, 0	1, 1	–	–	$\surd$

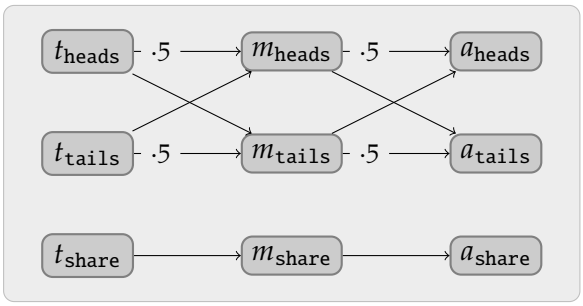
Figure 1.9: Matching pennies with cooperation option

But now, consider a slightly adapted version of this game. Suppose that while Bob is out of the room, either the coin is flipped or the judge tells Alice that it's "cooperation time." If it is officially cooperation time, and Bob guesses correctly that it is, both Alice and Bob win. But if Bob guesses on a coin flip outcome although it is actually cooperation time (or vice versa), then both Alice and Bob lose. Suppose, moreover, that Alice can now additionally announce that it is cooperation time whenever she wants to without constraints as to truth. The resulting game is given in figure 1.9.

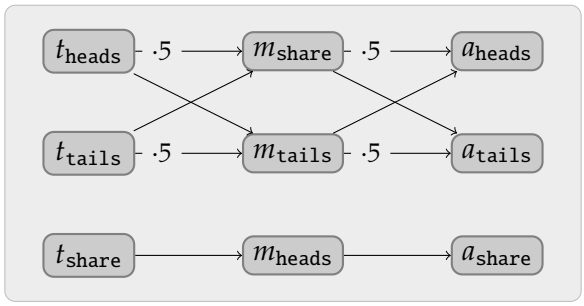
Ask yourself now, what you consider a natural way of playing this game (for the first time and only once). To my mind, it is absolutely natural to expect that Alice will use $m_{\texttt{coop}}$ if it is cooperation time and that she will use whatever other message, but certainly not $m_{\texttt{coop}}$, if it is not cooperation time; Bob, on the other hand, I would clearly expect to trust and believe message $m_{\texttt{coop}}$ and I would also expect him not to believe that either $m_{\texttt{heads}}$ or $m_{\texttt{tails}}$ was sent if it is cooperation time. Technically speaking, this comes down to saying that I believe that the equilibrium in figure 1.10a is more natural than that in figure 1.10b, although both are PBEs of the cheap talk game in figure 1.9.

If you share this judgement, you basically have an intuition about the effect of conventional meaning in a game where false signaling is possible and not penalized; you have an intuition about *credibility of messages* in a cheap-talk signaling game. The abstract perspicuity of the stylized example should not obscure the appreciation of a conceptually very important point: whether semantic content is to be taken seriously depends on the particular constellation of preferences of interlocutors; it is a matter of rational deliberation based on details of the context of utterance —a pragmatic inference if you wish to call it so— whether or not to believe certain semantic information.

It also does not matter that the above example is too abstract and too precise to faithfully match most of our everyday conversations. The point is simply that there are situations, even if marginal, that make it absolutely clear that it is our intuitions about *rational* language use that delineate which



(a) Intuitive equilibrium



(b) Crooked equilibrium

Figure 1.10: Some equilibria of the extended matching pennies game

part of semantic meaning is to be ignored and which is to be taken seriously under a given strategic constellation. But that means that semantic meaning should (somehow) be implemented in the solution concept, not the context model. It would simply be absurd to try to account for our intuitions about credible cheap talk by restricting the sender's strategy space or imposing a cost on sending certain messages in certain situations. We should ideally let our rational agents, not the modeller, decide what to say and what to believe.

Credibility-Based Refinements

In fact, game theorists have addressed this issue, and have asked precisely under which formal conditions a message is credible in a cheap talk game (see Farrell and Rabin 1996, for overview). There are strictly speaking two prominent contexts in which game theorists ask for the credibility of messages.[24]

24. There are other aspects of the notion of credibility that have been studied in game theory. One prominent line of conceptual analysis which is *not* explicitly dealt with in this thesis addresses the concept of *speaker credibility* (cf. Sobel 1985): is the person I am talking to reliable and trustworthy; have her past actions convinced me of her integrity? In the present context, such issues of long-term reputation and personal history between players are not

One is in (cheap talk) pre-play communication (cf. Farrell 1988; Rabin 1994) where players state which actions they intend to perform during a play of the game. If such statements are assumed to be non-binding, the issue of credibility is pressing and takes the form of asking when a signal is *self-committing*: a signal "I will play such and such" is SELF-COMMITTING if, roughly, it creates an incentive for the speaker to fulfill it (cf. Farrell and Rabin 1996, p. 111). In contrast to pre-play announcements of intentions, we are presently interested in whether a message in a signaling game is *self-signaling*: a message is SELF-SIGNALING if, roughly, its utterance is sufficient evidence that it is true. Self-signaling messages should thus be believed, and it is in this sense that we speak of credibility of cheap talk in signaling games.

Which messages are intuitively credible in a given case depends on several aspects of the strategic situation. To begin with, whether a message is credible or not obviously depends on *its* semantic meaning, but also on the set of other available messages and *their* semantic meaning. Moreover, of course, the agents' utilities, in particular the degree of preference alignment in various states, will also play a crucial role. Without going into any detail, it is palpable that a satisfactory definition of message credibility is not too easy to come up with. Still, this is what game theorists have tried in order to refine basic solution concepts such as rationalizability (Rabin 1990; Zapater 1997) or equilibrium (Myerson 1989; Farrell 1993; Matthews et al. 1991). The general idea behind such credibility-based refinements is basically to define in the abstract when exactly a message is credible, and then to require that all credible messages be treated as such by the solution concept.

The approach presented in this thesis is the reverse. Instead of defining a notion of credibility and deriving a refined solution concept, I suggest to refine the solution concept and derive a notion of credibility. My solution concept —to be spelled out in the subsequent chapter— implements semantic meaning as a reasoning bias of agents. This, as it turns out, not only solves the problem of equilibrium selection in relevant pragmatic applications, but also yields a novel and simple notion of message credibility in the abstract.

TOWARDS A SOLUTION CONCEPT AS A PRAGMATIC THEORY. To sum up at this point, GTP shares a problem with other applications of game theory, namely the need to specify an appropriate solution concept that *uniquely* yields the intuitively/empirically desirable predictions. There does not appear to be any

addressed explicitly. To the extent that such matters play a role in a given situation, we have to imagine them expressed in the utility functions of a given signaling game.

established game theoretic notion that gives satisfactory predictions in the pragmatic realm. However, this lacuna is perhaps more chance than doom, because it leaves us with the freedom to define a feasible solution concept based on exactly those assumptions —preferably independently and empirically motivated— about human behavior and cognition that we deem relevant in natural language use and interpretation. There is no reason why we need to stick to traditional concepts of equilibrium, for instance. Empirical results of experimental game theory and psycholinguistics should ideally inform the formalization of both context models and solution concept. Empirical research in game theory is blooming (see Camerer 2003), and empirically informed applications of game theory to pragmatics should —and are beginning to— follow suit (see Sally 2003; de Jaegher et al. 2008). In particular, an epistemic approach to game theory seems like a very promising platform to formally implement empirically motivated assumptions about the psychology of reasoners. Consequently, the next chapter offers a novel solution concept, spelled out in terms of epistemic assumptions about reasoning agents, which specifically models psychologically biased and possibly resource-limited reasoning about natural language.

Chapter 2

The Iterated Best Response Model

> "At first I basically thought: What the fuck? And then I thought: You've got to be kidding me. And then I began to sort of think, Oh no." (Gessen 2008, p. 15)

> "[T]he natural way of looking at game situations [...] is not based on circular concepts, but rather on a step-by-step reasoning procedure." (Selten 1998, p. 421)

Chapter Contents

2.1 · Focal Points & Iterated Best Response · 44

2.2 · The Vanilla Model · 53

2.3 · Forward Induction · 76

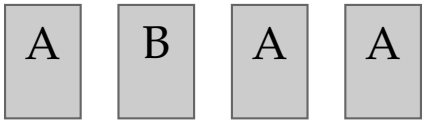
2.4 · Overview and Comparison · 89

2.5 · Semantic Meaning and Credibility · 107

This chapter presents a model of *iterated best response reasoning with focal points* (the IBR model). The chapter is structured as follows. I will first try to motivate the core assumptions of my approach in section 2.1 by reviewing key results of behavioral game theory. Then I will spell out a plain version of the IBR model in section 2.2 and apply it to some illustrating examples. The basic model will be supplemented with an additional refinement: section 2.3 introduces *forward induction reasoning*. Section 2.4 reflects back on the proposed model and its variations, and compares the lot to other relevant models in pragmatics and game theory. Finally, section 2.5 comes back to the question whether the IBR model really properly implements conventional meaning in a pragmatic solution concept, and compares the present approach with previous approaches to message credibility.

2.1 Focal Points & Iterated Best Response

As a means of introducing in rough outline iterated best response reasoning with focal points, consider the following simple 'hide-and-seek' game. There are four labelled and linearly arranged doors, as shown here:



One player, called *Hider*, hides a prize behind any of these doors and a second player, *Seeker*, simultaneously guesses a door. *Seeker* wins iff *Hider* loses iff *Seeker* chooses the door where *Hider* hid the prize. The payoff structure for this game is the following (*Hider* is the row player):

	Door 1	Door 2	Door 3	Door 4
Door 1	0,1	1,0	1,0	1,0
Door 2	1,0	0,1	1,0	1,0
Door 3	1,0	1,0	0,1	1,0
Door 4	1,0	1,0	1,0	0,1

When looking at the game in this abstract form, there is nothing that should prejudice any of the four doors over any other for either *Hider* or *Seeker*. There is exactly one unique mixed Nash equilibrium in this strategic game: both players choose a door completely at random with probability $1/4$ for each

door.[1] However, the different labeling of doors and their linear arrangement does seem to make a difference to human reasoners. There are, as the behavioral game-theorist would say, NON-NEUTRAL PSYCHOLOGICAL FRAMING EFFECTS in the way the game is presented. And, indeed, when Rubinstein et al. (1996) put this condition to the test, they found that the following percentage of subjects chose the various doors:

	A	B	A	A
Hider	9%	36%	40%	15%
Seeker	13%	31%	45%	11%

This result deviates significantly from a flat 25% choice of every door that we would expect if reasoners played the unique mixed Nash equilibrium. Something in the presentation of the game, the labelling of doors and their linear arrangement, must have prejudiced human reasoners to consider some alternatives more salient than others. This is also highly plausible by introspection: the door labeled B very obviously sticks out, and similarly so do the left- and right-most doors.

Experiments following this paradigm have been multiply replicated. Surveying these, Crawford and Iriberri (2007) argue that Rubinstein et al.'s empirical results in this and similar 'hide-and-seek' games on non-neutral landscapes can best be explained by an ITERATED BEST RESPONSE MODEL WITH FOCAL STARTING POINTS, which I will henceforth call an IBR model (with focal points) for short.[2] Such a model, basically rests on two assumptions, namely that:

1. there are FOCAL POINTS in the presentation of the game that attract the attention of reasoners *before* they engage in further strategic deliberation; and that

2. starting from this initial focal prejudice of attention, players use ITERATED BEST RESPONSE REASONING at different levels of strategic sophisti-

1. Strictly speaking, the notion of a *mixed* Nash equilibrium of a strategic game has not been introduced explicitly, but it is also not essential here and in the following. The interested reader is referred to the standard textbooks.

2. To say here that the IBR model explains the data *best* needs a careful hedge, if we want to be precise and fair. Crawford and Iriberri (2007) show that their IBR model with focal point reasoning provides the best model from a set of competing alternative models if several factors are taken into account: generality and portability of the model, theoretical parsimony, *and* econometric fit of the data. Looking at econometric fit of the data alone, the IBR model of Crawford and Iriberri does not do better than some of the alternatives, but also not worse.

> cation, i.e., they compute best responses to focal point behavior, to which they compute a best response (if they can), to which they compute a best response (if they can), and so on.

For example, according to an IBR model with focal points for the above 'hide-and-seek' game *Hider* might reason as follows.[3] *Hider* might start her deliberation with the focal point, saying to herself: "Obviously, the door labeled B sticks out," and then go on reasoning about *Seeker*'s behavior based on this: "So, I expect that if *Seeker* doesn't think much about what he's doing, he will choose this door." This is then where *Hider* would anticipate the behavior of a naïve, unstrategic player. Based on this, *Hider* would act rationally by thinking: "But then, I should *not* hide the prize there, but choose another door." But *Hider* may also anticipate that *Seeker* may anticipate her own best response; *Hider* may think: "But, hey, if *Seeker* thinks the same, I probably should hide the prize exactly behind door B." Clearly, for a zero-sum game this reasoning pattern will soon start to loop. (We will come back to this feature in section 2.2.1.) For the time being, the point of interest is that to assume that subjects perform roughly this kind of reasoning explains well the empirical data of Rubinstein et al. Let me therefore enlarge briefly on both assumptions, focal points and iterated best response reasoning, in order to motivate their respective and conjoined use as a model of pragmatic reasoning.

2.1.1 Semantic Meaning as a Focal Point

SCHELLING POINTS IN COORDINATION. The idea of focal points, that somehow attract our attention and therefore psychologically bias our reasoning patterns, is very natural. It is also familiar, in slightly different form, from Thomas Schelling's ground-breaking work on equilibrium selection in strategic coordination games (Schelling 1960). Schelling's idea was that independent coordination choices will often converge on the most *salient* option. For example, if two people have to independently make a choice such as where to go meet the other person somewhere in New York city when they cannot communicate a meeting place beforehand but know that it's commonly known to both sides that they are facing exactly this coordination problem,

3. This exposition is simplified, assuming that only the door labeled B is focal. The interaction of two focal points of possibly differing strengths of attraction further complicates the example, but I will gloss over this here because I merely want to introduce the general idea of IBR reasoning.

then it is not only fairly natural to assume, but also empirically supported (see Camerer 2003, chapter 7 for overview) that people will coordinate on choices that are somehow psychologically salient, such as, in the present example, Grand Central Terminal in Midtown Manhattan. What counts as salient under which circumstances is a separate, interesting, but ultimately empirical question.[4] Nonetheless, Schelling's insight remains: people are guided by psychological salience when choosing among several possible coordination equilibria that are, as far as utilities are concerned, equally good.

Schelling's idea of focality has had prominent influence on some applications of game theory to linguistics and philosophy of language. Lewis' analysis of the notion of 'convention' in terms of signaling games is inspired by Schelling's insight that precedence may act as a focal element (Lewis 1969). Parikh, on the other hand, motivates his use of Pareto-dominance as a second-order selection principle on sets of equilibria with reference to focality of Pareto-efficiency (Parikh 2001).

Focality as Starting Point of Deliberation. The role of focal points in an IBR model is slightly different though. First of all, focality in the IBR model is not a second-order selection criterion on top of standard equilibrium notions. Similarly, whereas Schelling's focal points are what most people would expect to be a commonly expected coordination point, focal points in the IBR model are not —in a manner of speaking— the *outcome* of reasoning about a game situation but rather the *starting point*. Focal point reasoning in IBR is also not confined to coordination games, as the above 'hide-and-seek' game illustrates: reasoners might convince themselves that playing a focal strategy is *not* a good idea. In other words, reasoners may reason themselves away from focality, rather than being attracted by it through or after deliberation. Still, the general idea of a psychologically attractive option that most if not all people will notice and know that most if not all people will notice etc., is the same.

Semantic Meaning as Focal. Focal point reasoning, I would like to suggest, is fairly intuitive also for models of natural language interpretation. The model of pragmatic reasoning that this chapter spells out therefore rests on the following Focal Meaning Assumption: semantic meaning is focal in the sense that pragmatic deliberation —to be identified as a sequence of

4. Schelling wrote: "One cannot, without empirical evidence, deduce what understandings can be perceived in a nonzero-sum game of maneuver any more than one can prove, by purely formal deduction, that a particular joke is bound to be funny." (Schelling 1960, p. 164)

best responses— departs from semantic meaning as a psychological attraction point of interlocutors' attention. In other words, the semantic meaning of messages is a focal point, I would propose, much like the door labeled B in the above 'hide-and-seek' game: even though strategically semantic meaning is *not* binding, it is fairly intuitive to start pondering how to use an expression and what might have been meant by its use by assessing first the expression's semantic meaning.

Thus conceived, the Focal Meaning Assumption is a solution to a technical problem —the problem how to implement semantic meaning non-bindingly in the solution concept for games— which has a general, independent empirical motivation in the psychology of reasoners. On top of that, I believe that the Focal Meaning Assumption is not entirely implausible for the intended purpose either: it is not unnatural to assume that the conventional meaning of an expression provides the best first clue to utterance meaning. Using rationalistic vocabulary we could say that, given a semantically meaningful message, the hearer would like to rationalize why the speaker said what he said. So, as a starting point of his deliberation the hearer asks himself, what he would do if the message was indeed true. But then he might realize that the sender could anticipate this response. In that case, the hearer is best advised to take the sender's strategic incentives —her preferences and action alternatives— into consideration. Similarly, a naïve sender might just say whatever is true at a given occasion. But with some more pragmatic sophistication she might reason her way up the IBR ladder where she includes her expectations of the receiver's responses to her naïve sending strategy. The resulting hypothetical reasoning on both the sender and the hearer side can be modelled as a sequence of iterated best responses that, crucially, takes its origin in a focal point constituted by semantic meaning. And this is, in bare outline, the solution concept that this chapter will put forward.

2.1.2 Iterated Best Response Reasoning as Pragmatic Inference

IBR models not only help implement possible psychological reasoning biases in the form of focal starting points of the deliberation, but they also intend to capture (some of) the natural resource-bounded limitations of actual human reasoning. That human reasoning is bound, in a manner of speaking, to a finite, even narrow horizon of strategic sophistication has been demonstrated repeatedly in multiple laboratory experiments on strategic reasoning. In simplified terms, the upshot of this empirical research is that although nearly all

subjects behave rationally in laboratory games, far fewer subjects trust the rationality of others when making their own decisions, even fewer people are certain of others' belief in others' rationality, and so on. What is generally at stake here can be appreciated also intuitively based on so-called 'p-beauty contest' games, which I will tend to presently. A subsequent look at subjects' reasoning in particular dynamic games will help refine the general picture.

'p-Beauty Contest' Games. For a start, let's have a look at experiments on so-called 'p-beauty contest' games, which have been tested extensively (see Nagel 1995; Ho et al. 1998; Camerer 2003; Camerer et al. 2004).[5] In such a 'p-beauty contest' each player from a group of size $n > 2$ chooses a number from zero to 100. The player closest to p times the average wins. When this game is played with parameter $p = {}^2/_3$ by a group of subjects who have never played the game before, the usual group average lies somewhere between 20 and 30, which is curiously far from the group average zero which we would expect from common (true) belief in rationality and which is the only Nash equilibrium in this game.

The reasoning with which a player may arrive at the conclusion that zero is the analytically best choice in a 'p-beauty contest' with $p = {}^2/_3$ is the following: a rational player will not play a number bigger than 67 because any such number has a lower chance of winning than exactly 67;[6] but if a rational player believes that her opponents are rational and will therefore realize this much, she should not play any number higher than ${}^2/_3 \times 66 = 44$ by the same reasoning; again, a rational player who is convinced that her opponents are rational and believe in the rationality of others will play maximally ${}^2/_3 \times 44 \approx 29$, and so on. Further iterated steps of such reasoning will lead to

5. The name of these games, however, is slightly misleading. It originates in an observation by John Maynard Keynes who likened stock markets to a newspaper contest in which readers were encouraged to guess which face most readers would choose as the most beautiful (Keynes 1936). The newspaper's beauty contest is actually a coordination game much in the sense of looking for a Schellingesque focal point in a coordination game: guessers need to guess what others guess (what others guess etc.) to be the most beautiful face. In 'p-beauty contest' games, the element of coordination is broken in favor of a more abstract and revealing game design.

6. The notion of rationality at stake here is that no action will be played which is stochastically dominated: an action A is stochastically dominated by an action B if the chance of ascertaining a fixed amount of payoff when playing B is higher than when playing A. The process that is outlined informally here is one of iteratively removing stochastically dominated actions. I am glossing over interesting technical detail here for the sake of exposition.

the conclusion that zero is the best choice, in fact the only choice compatible with common belief in rationality.

Nonetheless, but perhaps unsurprisingly, few subjects in experimental plays of this game choose zero — be that the first time they play this game or in later rounds after having observed the behavior of other players. Rather, a large pool of data on these kinds of games suggests that "the typical subject uses only one or two steps of reasoning (starting from 50)" (Camerer 2003, p. 218).

A non-zero choice is by no means a bad choice, of course. If (you believe that) everybody else chooses relatively high numbers, you do not want to choose too low a number yourself. What a number choice in this game actually represents is a player's estimate of the estimate (of the estimate ...) of other players: provided a player is rational, her choice will be around $^2/_3$ her estimated average; but that means that any choice of number other than zero is either irrational —which we will exclude— or indicative of a belief that some or most other players are a little less smart than the choosing player herself. What any choice other than zero therefore truly expresses is a belief in the relative reasoning *in*capabilities of others and a certain amount of OVERCONFIDENCE: bluntly put, a non-zero choice says "I think I am smarter than you guys are (on average)."

In sum, 'p-beauty contest' games show, both experimentally and intuitively, a general healthy tendency of human subjects to distrust other players' ideal and flawless rational behavior and/or reasoning capabilities. Similar results have been obtained from experiments on different kinds of games. Stahl and Wilson (1995), for instance, tested subjects' performance on static games some of which had unique solutions in either one or two steps of iterated strict dominance. Camerer (2003) discusses a wealth of experiments on similarly 'dominance-solvable' games, amongst others variants of the centipede game, variants of the muddy children puzzle and the 'electronic mail' game (Rubinstein 1989). Despite the heterogeneity of the tested games, a rough and general conclusion —to be scrutinized presently— seems feasible: subjects' performance in experimental conditions drops to the extent that the tested choice requires higher levels of iterated reasoning. The more careful question we should ask though is: what exactly is the limitation in subjects' game theoretic reasoning due to?

ITERATED DOMINANCE IN DYNAMIC GAMES. A dynamic two-player game like that in figure 2.1 may help shed light on this issue. In this game, first player

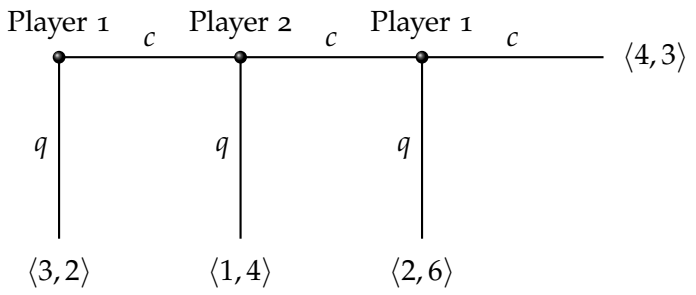


Figure 2.1: A dominance-solvable dynamic game

1 makes a choice to continue (c) or quit (q), then player 2 does the same if player 1 continued to play initially, and finally player 1 gets another choice to continue or quit if player 2 decided to continue too. Clearly, if ever player 1's last choice point is reached she should choose to continue because that will give her a payoff of 4 as opposed to 2. Then, if player 2 realizes that player 1 will play c at her last choice point, player 2 should quit when she gets the chance, securing a payoff of 4 instead of 3. But then, anticipating player 2's behavior, player 1 should already initially quit the game to obtain 3 utils instead of only 1 util.

This kind of BACKWARD INDUCTION reasoning on player 1's part corresponds with her ascribing to player 2 both rationality and the belief that player 1 will make a rational choice at her last choice point (Aumann 1995; Stalnaker 1998). In other words, player 1 will play the prediction of backward induction, if she believes that (i) player 2 is rational and that (ii) player 2 believes that player 1 herself is rational (assuming, that is, that the game structure is common knowledge).

This suggests that player 1's initial choice, whether to continue or quit, is diagnostic of the depth of strategic reasoning that player 1 is capable of, including the depth of strategic reasoning that player 1 is able to ascribe to player 2. Hedden and Zhang (2002) indeed argue that player 1's initial choice is indicative in particular of whether player 1 is using first- or second-order THEORY OF MIND REASONING (ToM reasoning) (Premack and Woodruff 1978): according to common classification, zeroth-order ToM reasoning is reasoning that takes into account one's own desires and beliefs about the state of the world only, first-order ToM reasoning takes into account others' desires and (zeroth-order) beliefs, and $(n+1)$-th-order ToM reasoning takes into account others' desires and n-th-order beliefs. According to Hedden and Zhang, player 1's initial choice to continue is indicative of her using first-order theory

of mind:[7] if player 1 does not reason about player 2's first-order theory, she must think that player 2 is a zeroth-level reasoner who does not take player 1's incentives into account, and therefore, Hedden and Zhang assume, a first-order player 1 will assume that player 2 will choose to continue, hoping for an average payoff of $4\frac{1}{2}$. On the other hand, if player 1 uses second-order ToM reasoning she will choose to quit initially, as predicted by backward induction and its above epistemic justification.

Hedden and Zhang conclude based on their experimental data that most subjects initially only apply first-order ToM reasoning, and only later possibly advance to second-order ToM reasoning when playing against a first-order confederate player 2. It is not essential to discuss the soundness of this conclusion in minute detail. It suffices to note that the conclusion as such is in line with the bulk of research on dominance-solvable games: subjects are capable of one or two levels of iterated reasoning. Hedden and Zhang then suggest that this may be due to a lack of ToM reasoning capability. For clarity, this is a thesis opposed to the idea that subjects, for instance, lack *trust* in rationality. To say, as Hedden and Zhang do, that subjects lack the *conceptual grasp* or the *computational resources* necessary for higher-order ToM reasoning is different from saying that belief in (belief in...) rationality is waning proportional to the depth of nesting of belief in rationality. Which position is correct, or whether even both are or neither is, is a matter for more refined empirical research that this thesis does not contribute to, unfortunately. Although the model I will present in this chapter is, as far as I can see, compatible with both explanations of human reasoning limitations in dominance solvable games, I will adopt Hedden and Zhang's idea that it is general ToM reasoning that is difficult —either to grasp or perform— and *not* so much faith in rationality that is lacking.[8]

2.1.3 Strategic-Type Models

In order to capture the idea that human reasoning is at the same time bounded in the number of analytical steps and overconfident in assuming that others can be outperformed, several behavioral game theorists have postulated mod-

7. Several things that are crucial to evaluating Hedden and Zhang's experimental design are left out in this short exposition. The interested reader is referred to the original paper, as well as Colman (2003) and Zhang and Hedden (2003).

8. I adopt this position here partly also because it makes for a neater model, but I will come back to a closer discussion of the role of ToM reasoning in chapter 4, where I discuss the connection of the IBR model with bidirectional optimality theory.

els featuring different *strategic types* of players (Stahl 1993; Stahl and Wilson 1995; Holt 1999; Crawford 2003; Camerer et al. 2004). Detailed differences in models notwithstanding, a strategic type captures the level of sophistication of a player and corresponds to the number of steps that the agent is able to (and/or in fact does) compute in a sequence of iterated best responses. This number of steps is bounded above by the maximal order of ToM reasoning that the agent is capable of. It is then the set of *all* such strategic types, with their beliefs and behavior, that forms the prediction of a strategic-type model.

Such strategic-type models are good predictors of experimental data, because they are often simpler than competing theories (involving fewer parameters), and more generally applicable at equal econometric fit (see Camerer 2003; Camerer et al. 2004; Crawford and Iriberri 2007). Additionally, these models are also conceptually appealing for several reasons. First of all, these models allow the implementation of focal points in a natural manner as starting points of strategic reasoning. Moreover, strategic type models take seriously the natural resource boundedness of ToM reasoning, as demonstrated in the last section. The advantage of this is that a strategic type model, as an analytic solution concept, also yields predictions about boundedly rational behavior that possibly falls short of the classical game theoretic ideals of equilibrium or (play consistent with) common belief in rationality.

In the following section I will propose a model of strategic types of senders and receivers in signaling games. The assumption that semantic meaning is focal cashes out in the stipulation of level-zero players that do not engage in strategic reasoning at all: they are blind to their opponent's strategy and preferences, and only take into account the semantic meaning of messages. In particular, a level-zero sender would like to send arbitrarily any message that is true; similarly a level-zero receiver would simply believe all messages literally. Given a specification of level-zero players, we can define the behavior of level-$(k+1)$ players by induction. A level-$(k+1)$ player believes that his opponent is a level-k player and will play a best response to this belief.

2.2 The Vanilla Model

Recall from the previous chapter that a signaling game (with meaningful signals) is a tuple

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

where sender S and receiver R are the players of the game; T is a set of states; $\Pr \in \Delta(T)$ is a probability distribution over T with full support; M

is a set of messages that the sender can send; $[\![\cdot]\!] : M \to \mathcal{P}(T)$ gives the semantic meaning of a message; A is the set of receiver actions; and $U_{S,R} : T \times M \times A \to \mathbb{R}$ are utility functions for both sender and receiver.

Generally speaking, the IBR model proposed here defines strategic types of players in terms of their beliefs about the opponent's behavior. More concretely, a level-$(k+1)$ player believes that she is playing against a rational level-k opponent. That is to say that I will assume here that each higher level player believes that she is *exactly* one level more sophisticated than her opponent.[9] Additionally to that we may allow level-$(k+1)$ players to have further prejudices and beliefs about the belief formation and behavior of their opponents. For ease of exposition though, I will first spell out a vanilla version of the IBR model without such extra assumptions.

2.2.1 Strategic Types and the IBR Sequence

Before plunging into the definitions of strategic types, a word of caution is in order. Since the IBR model defines player types in terms of beliefs about opponent behavior, the notation I will use is strictly —but harmlessly— ambiguous: S_k, for instance, will denote both (i) a sender of strategic level k as an abstract entity defined by the IBR model, but also (ii) the set of pure strategies representing the *unbiased belief* (see below) of R_{k+1} in his opponent's behavior. Analogously, of course, for R_k and S_{k+1}.

Level-Zero Players. The beginning of the IBR sequence is defined by types who adhere strongly to 'semantics only' in line with the Focal Meaning Assumption argued for above. I will assume that S_0 plays an arbitrary truthful sender strategy. A pure sender strategy s is TRUTHFUL iff $t \in [\![s(t)]\!]$ for all t. Hence, let S_0 be the set of all truthful sender strategies:

$$S_0 = \{s \in \mathsf{S} \mid \forall t \in T : \ t \in [\![s(t)]\!]\} .$$

Additionally, I will assume that R_0 plays an arbitrary strategy that is rational given a literal, semantic interpretation of the receiver message. A LITERAL INTERPRETATION is a posterior belief $\mu_0(\cdot|m) = \Pr(\cdot|\,[\![m]\!])$ which results from updating the prior beliefs with the semantic meaning of the observed message. In general, if $\delta \in \Delta(X)$ is a probability distribution over set X, then

9. This simplifying assumption makes the model more tractable and enables easy application for our linguistic purposes, but it is also unrealistic in several respects. Other models have made other design choices in the definition of higher level types, and I will come back to a thorough discussion and comparison of models in section 2.4.

the conditional probability of event $Y \subseteq X$ conditional on event $Z \subseteq X$ is calculated by Bayesian update:

$$\delta(Y|Z) = \frac{\delta(Y \cap Z)}{\delta(Z)}.$$

For a level-zero receiver, literal interpretation is such a Bayesian update of his priors with the event that the observed message is true. In this sense, a level-zero receiver considers the semantic meaning of an observed message, but does not take his opponent's strategy into account. Let $R_0 = \text{BR}(\mu_0)$ be the set of all rational responses to a literal interpretation.

Unbiased Beliefs. In the vanilla version of the ibr model, player types of level $k+1$ are simply defined as best responding to *unbiased beliefs* that their opponent is a level-k player, without any further restrictions on these beliefs. An unbiased belief in some finite set X is the belief that all $x \in X$ are equally likely and that all $y \notin X$ have probability zero. To use *unbiased* beliefs about possible opponent behavior is to average over any possible hunch or conjecture an agent may have about her opponent's behavior and to apply the 'principle of insufficient reason' as a strict tie-break rule at every iteration step. We may think of this as essentially a simplifying assumption that keeps the mathematics simple and allows for more straightforward computation in linguistic applications. We will come back, though, to a more thorough conceptual characterization of this assumption in section 2.4 of this chapter.

Notice that, for instance, an unbiased sender belief that the receiver is playing a strategy in some set $\mathsf{R}' \subseteq \mathsf{R}$ is entirely defined by the set R' itself. I will therefore use 'loose typing' and take R' to refer to either a set of pure receiver strategies or the corresponding sender belief, which strictly speaking should be represented as a behavioral strategy. The same applies, mutatis mutandis, to the receiver's unbiased beliefs in a set of sender strategies.

Higher Level Types. With this notational convention, the definition of the ibr sequence becomes very simple: S_{k+1} has an unbiased belief that she is facing R_k, so —with full use of loose typing— R_k simply *is* S_{k+1}'s belief of her opponents behavior. Her own rational behavior is then defined as:

$$S_{k+1} = \text{BR}(R_k).$$

For the receiver the situation is only slightly more complicated. If there are surprise messages under the belief S_k, then an unbiased receiver belief in a

set of sender strategies S_k does not necessarily yield a single unique posterior μ consistent with S_k to which the receiver could best respond.[10] The vanilla IBR model is simply unrestricted here and says that R_{k+1} will adopt *any* posterior μ which is consistent with the belief S_k. So, formally, let $\Pi_{R_{k+1}}$ be the set of all triples $\langle \Pr, S_k, \mu \rangle$ such that μ is consistent with S_k and define:

$$R_{k+1} = \mathrm{BR}(\Pi_{R_{k+1}}).$$

The vanilla IBR model consequently predicts that if m is a surprise message given belief in S_k, then R_{k+1} will respond to m with *any* action that is *zero-order rationalizable* in the play after m, i.e., rational for *some* belief $\delta \in \Delta(T)$ about which state is actual. Put formally, the set of ZERO-ORDER RATIONALIZABLE ACTIONS $A^*(m) \subseteq A$ after observing message m are all actions rational under some belief in $\mu(\cdot|m) \in \Delta(T)$:

$$A^*(m) = \left\{ a \in A \mid \exists \mu \in (\Delta(T))^M \; a \in \arg\max_{a' \in A} \mathrm{EU}_R(a', m, \mu) \right\}.$$

Clearly, if m is a surprise message, the vanilla IBR model yields

$$R_{k+1}(m) = A^*(m).$$

LIMIT PREDICTION. The IBR model defines an infinite sequence of ever more sophisticated players. It is important to stress that the IBR model is *not* an equilibrium solution concept, and that the behavior of boundedly rational types belongs to its predictions even though such behavior may fall short of the idealized predictions under common belief in rationality.

Nonetheless, the IBR model also makes predictions about unbounded ToM reasoners, so to speak. To see this, notice first of all that for finite T and M the IBR sequence always enters a cycle after some $k \in \mathbb{N}$: since there are only finitely many pure sender strategies for finite sets T and M, there are also only finitely many sets of such strategies; and since R_{k+1} is completely determined for a given S_k, the IBR sequence is bound to repeat itself. In a sense, we could consider the *limit behavior* of the IBR sequence, i.e., the set of all pure sender and receiver strategies that are repeated infinitely often, as the model's abstract prediction of *idealized pragmatic reasoning*. Let me elaborate on this idea.

10. Recall that surprise messages are messages that are not expected to be used given a belief about the sender's behavior, in this case given the belief S_k. We will come back to (the sender's beliefs about) the receiver's interpretation of surprise messages later in this chapter, in section 2.3.

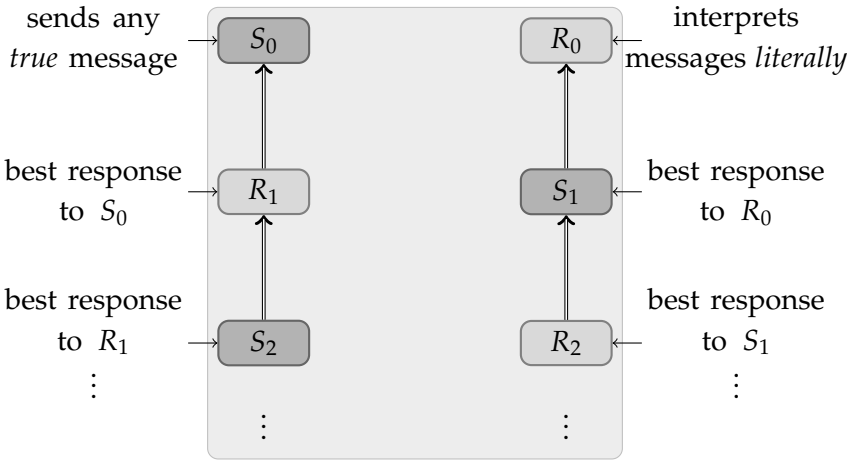


Figure 2.2: Schema of the IBR-sequence

The IBR model presented here defines two strands of iterated best response reasoning, one starting with a naïve sender and one starting with a naïve receiver. What results is a picture of IBR reasoning as schematized in figure 2.2: two separate strands of iterated best responses. We could speak of the S_0-sequence and the R_0-sequence respectively. It is relatively easy to see on an intuitive basis that each strategic type in such a double IBR sequence represents, in a sense, a certain resource limitation: not every strategic type necessarily has beliefs that are consistent with common belief in rationality; some strategic types' beliefs might only be compatible with some finite approximation of nestings in belief in rationality.

In particular, the picture is the following. The S_0-sequence contains all odd receiver types and all even sender types. The R_0-sequence contains all even receiver types and all odd sender types. As for the S_0-sequence, level-zero senders are possibly irrational, but level-1 receivers are rational, but need not believe that their opponents are rational; in turn, level-2 senders are rational and believe that their opponents are rational but not necessarily that their opponents believe in rationality etc.; this generalizes to saying that any level-$(2k-1)$ receiver is rational and believes in at least $2k-2$ nestings of belief in rationality and that a level-$(2k)$ sender is rational and believes in at least $2k-1$ nestings of belief in rationality ($k > 0$, of course). A similar fact holds for the R_0-sequence: level-$(2k-1)$ senders are rational and believe in at least $2k-1$ nestings of belief in rationality, while level-$(2k)$ receivers are also rational and believe in at least $2k$ nestings of belief in rationality.

That the IBR model represents cognitively limited, possibly too limited,

reasoners is crucial for explaining natural shortcomings of pragmatic reasoning and the development of pragmatic reasoning competence. Nevertheless, the IBR model also makes predictions, in a sense, about unrestricted, resource-unbounded IBR reasoners: since for finite sets T and M the IBR sequence will cycle, there are strategic types, even for finite levels k, whose beliefs and behavior is compatible with common belief in rationality: *any* strategic type that occurs in a cycle is repeated infinitely many times and therefore compatible with an unbounded nesting of belief in rationality. It is for this reason that I will speak of the set of sender and receiver types that are repeated infinitely often as *the model's prediction of idealized pragmatic reasoning in the limit* or for short the model's LIMIT PREDICTION. I will use the notation

$$\begin{aligned} \mathsf{S}^* &= \{s \in \mathsf{S} \mid \forall i \, \exists j > i \, : \, s \in \mathsf{S}_j\} \\ \mathsf{R}^* &= \{r \in \mathsf{R} \mid \forall i \, \exists j > i \, : \, r \in \mathsf{R}_j\} \end{aligned}$$

to collect all infinitely repeated strategies. It is then the tuple $\langle S^*, R^* \rangle$ that can be regarded as the IBR model's idealized solution, to compare it with other game theoretic solution concepts.

We will discuss the properties of the IBR model as a solution concept in more detail especially in section 2.4. For the moment, suffice it to conclude the exposition of the basic IBR model with a simple, obvious but noteworthy result about the model's limit prediction in case either sequence —starting with S_0 or R_0— reaches a fixed point, i.e., a cycle of length 1. It is fairly trivial to show that any fixed point of the IBR model in which there are no surprise messages under S^* is a perfect Bayesian equilibrium.[11]

Proposition 2.2.1. If $\langle S^*, R^* \rangle$ is a fixed point of an IBR sequence such that there are no surprise messages under S^*, then $\langle S^*, R^*, \mu^* \rangle$ is a perfect Bayesian equilibrium where μ^* is the unique posterior consistent with S^*.

Proof. 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. □

11. Proposition 2.2.1 does not generalize to arbitrary fixed points, because if there are surprise messages under S^*, then R^* is defined as the union of all best responses to some consistent belief in S^*. But that does not necessarily mean that there is a single unique posterior consistent with S^*, that rationalizes all of R^*'s reactions to surprise messages *at the same time*. Peeking ahead, the result does generalize, however, on the class of interpretation games, defined in chapter 3.1, that are used primarily for linguistic applications in this thesis.

2.2.2 Examples: Scalar & M-Implicatures

The workings of the IBR model will become much clearer when calculating some simple examples. Let's therefore first have a look at the some-all game for scalar implicature calculation, and subsequently at how the model deals with M-implicatures.

Scalar Implicature

The some-all game for scalar implicature calculation is given in figure 1.3 on page 21. For the sake of the example, let us assume that the prior probability distribution is flat, i.e., that $p = 1/2$. It will transpire that the vanilla IBR model uniquely selects the desired equilibrium behavior in this case.

The behavior of level-zero players is straightforward. S_0 will send some true message in each state:[12] $S_0^{t_{\exists\neg\forall}}$ will send only $m_{\texttt{some}}$, while $S_0^{t_\forall}$ might send either $m_{\texttt{some}}$ or $m_{\texttt{all}}$. As an unbiased belief of R_1, S_0 can then be perspicuously represented as follows:

$$S_0 = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\texttt{some}} \\ t_\forall & \mapsto & m_{\texttt{some}}, m_{\texttt{all}} \end{array} \right\}.$$

For clarity, this means that a level-zero sender is expected (by R_1) to be indifferent between sending $m_{\texttt{some}}$ and $m_{\texttt{all}}$ in state $t_\forall$.

Similarly, R_0's posteriors are easily calculated by Bayesian update:

$\mu_0(t\|m)$	$t_{\exists\neg\forall}$	$t_\forall$
$m_{\texttt{some}}$	$1/2$	$1/2$
$m_{\texttt{all}}$	0	1

and the resulting set of pure receiver strategies $R_0 = \mathrm{BR}(\mu_0)$ again is straightforwardly represented as:

$$R_0 = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_{\exists\neg\forall}, t_\forall \\ m_{\texttt{all}} & \mapsto & t_\forall \end{array} \right\}.$$

This much is nothing out of the ordinary: S_0 and R_0 are unstrategic players who simply incorporate semantic meaning into their behavior in the most straightforward fashion. But already at the next level of iteration things start

12. I will write S_k^t for a sender of information type t and strategic type k.

to become interesting. A level-1 sender will show what we could call SCALAR IMPLICATURE BEHAVIOR:

$$S_1 = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\texttt{some}} \\ t_{\forall} & \mapsto & m_{\texttt{all}} \end{array} \right\}.$$

S_1 will send messages corresponding one-to-one to states, because this is the most optimal way of behaving under S_1's belief that messages are interpreted literally. To break this prediction down in more detail, in state $t_{\exists\neg\forall}$ the expected utility of sending message $m_{\texttt{some}}$ given belief R_0 is $1/2$, but that of sending message $m_{\texttt{all}}$ is zero. Moreover, the expected utility of sending message $m_{\texttt{some}}$ in state $t_{\forall}$ is $1/2$ for S_1, while that of sending $m_{\texttt{all}}$ is 1. Whence that the scalar implicature behavior of S_1 is the only rational behavior given her belief in literal interpretation. Nonetheless, S_1 does not believe that her message $m_{\texttt{some}}$ is going to be understood as uniquely expressing $t_{\exists\neg\forall}$. Thus, although S_1 shows scalar implicature *behavior*, she does not yet have SCALAR IMPLICATURE BELIEFS, as we could say.

To specify R_1's behavior, we first have to calculate his posterior beliefs μ_1 which should be consistent with his unbiased belief in S_0. The only non-trivial part of this calculation is the value of $\mu_1(\cdot|m_{\texttt{some}})$. Here is the calculation based on consistency:

$$\begin{aligned} \mu_1(t_{\exists\neg\forall}|m_{\texttt{some}}) &= \frac{\Pr(t_{\exists\neg\forall}) \times S_0(m_{\texttt{some}}|t_{\exists\neg\forall})}{\sum_{t'\in T} \Pr(t') \times S_0(m_{\texttt{some}}|t')} \\ &= \frac{1/2 \times 1}{1/2 \times 1 + 1/2 \times 1/2} \\ &= 2/3. \end{aligned}$$

The result means that a receiver who believes that states are equiprobable at the outset and believes in the sender's strategy S_0 will come to believe after hearing message $m_{\texttt{some}}$ —if his posteriors are consistent with these two beliefs— that it is twice as likely that the true state of the world is $t_{\exists\neg\forall}$ rather than $t_{\forall}$. I will elaborate on this feature of the consistency requirement, which some readers might find surprising, in section 2.2.3. For the time being, suffice it to note that the complete resulting posterior of R_1 is:

$\mu_1(t\|m)$	$t_{\exists\neg\forall}$	$t_{\forall}$
$m_{\texttt{some}}$	$2/3$	$1/3$
$m_{\texttt{all}}$	0	1

Based on these posterior beliefs, R_1 will also show *scalar implicature behavior*:

$$R_1 = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_{\exists\neg\forall} \\ m_{\texttt{all}} & \mapsto & t_{\forall} \end{array} \right\}.$$

This is because R_1 will maximize his expected utility after each observed message based on μ_1, and this entails in particular that after $m_{\exists\neg\forall}$ action $a_{\exists\neg\forall}$ is uniquely chosen. Nonetheless, at this stage of pragmatic sophistication the posterior μ_1 does *not* actually rule out that the message $m_{\exists\neg\forall}$ could have been sent in state $m_{\forall}$, i.e., R_1 does not yet have *scalar implicature beliefs*.[13]

Eventually, the IBR model predicts that for all player types of level $k \geq 2$ we get the same prediction:

$$S_k = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\texttt{some}} \\ t_{\forall} & \mapsto & m_{\texttt{all}} \end{array} \right\} \qquad R_k = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_{\exists\neg\forall} \\ m_{\texttt{all}} & \mapsto & t_{\forall} \end{array} \right\}$$

$$\mu_k = \left\{ \begin{array}{lcll} & & t_{\exists\neg\forall} & t_{\forall} \\ m_{\texttt{some}} & \mapsto & 1 & 0 \\ m_{\texttt{all}} & \mapsto & 0 & 1 \end{array} \right\}$$

Level-k agents not only show scalar implicature *behavior*, but also scalar implicature beliefs: they believe that their opponents will also show scalar implicature behavior. With this, the IBR sequence has reached a fixed point after two rounds of iteration; indeed the same fixed point for *both* sequences. This is then the unique limit prediction of the model for the scalar implicature game. By proposition 2.2.1, $\langle S^*, R^*, \mu^* \rangle$ is a perfect Bayesian equilibrium, and in fact the intuitively appropriate one in the set of PBEs for this game. Thus conceived, the IBR model solves the problem of equilibrium selection for this game that the previous chapter worked out as a central problem of GTP. Moreover, the IBR model does so by implementing semantic meaning not in the game model but in the solution concept, much as we wanted it to.

M-Implicatures

The IBR model also deals surprisingly well with M-implicatures. This is noteworthy in the light of the fact that M-implicatures turned out problematic for

13. The distinction between implicature *behavior* and implicature *beliefs* will indeed play an explanatory role later in this thesis: in section 4.4 for an explanation of the peculiar developmental pattern of acquisition of pragmatic competence in young children.

pretty much all standard solution concepts in game theory, and many non-standard solutions have been tried within the realm of classical game theory and beyond it (cf. Parikh 2001; van Rooij 2004*b*; de Jaegher 2008).

Remember from section 1.1.2 that we want to explain how an unmarked form (9a) is paired with an unmarked meaning (9b), while a marked form (10a) is paired with a marked meaning (10b).

(9a) Black Bart killed the sheriff.

(9b) $\rightsquigarrow$ Black Bart killed the sheriff in a stereotypical way.

(10a) Black Bart caused the sheriff to die.

(10b) $\rightsquigarrow$ Black Bart killed the sheriff in a non-stereotypical way.

Before heading into calculation of this example, it may be worthwhile mentioning that Horn's division of pragmatic labor is often considered a phenomenon about language organization and hence something that needs to be dealt with by a theory of diachronic language change. This may well be correct for the inferences associated with overly complex causative constructions wherever a lexicalized causative exists, as in the contrast between (9a) and (10a). However, there are also cases of M-implicatures like those in (14)–(17) that do call for a synchronic treatment.

(14) a. Sue smiled.

b. $\rightsquigarrow$ Sue smiled genuinely.

(15) a. The corners of Sue's lips turned slightly upwards.

b. $\rightsquigarrow$ Sue faked a smile.

(16) a. Mrs T sang 'Home Sweet Home.'

b. $\rightsquigarrow$ Mrs T sang a lovely song.

(17) a. Mrs T produced a series of sounds roughly corresponding to the score of 'Home Sweet Home.'

b. $\rightsquigarrow$ Mrs T sang very badly.

Context Model. The signaling game that models abstractly the basic features of these inferences is given in figure 2.3. The utilities listed in this figure are response utilities. We should additionally assume that the long message $m_{\texttt{lng}}$, which would correspond to (10a), incurs a slightly higher *cost* than the short message $m_{\texttt{shrt}}$, which corresponds to (9a). On top of that, we should assume that $p > 1/2$, i.e., that the normal state of affairs $t_{\texttt{norm}}$, which corresponds

	$\Pr(t)$	$t_{\texttt{norm}}$	$t_{\texttt{abn}}$	$m_{\texttt{shrt}}$	$m_{\texttt{lng}}$
$t_{\texttt{norm}}$	p	1,1	0,0	$\surd$	$\surd$
$t_{\texttt{abn}}$	$1-p$	0,0	1,1	$\surd$	$\surd$

Figure 2.3: A context model for M-implicatures

to (9b), is however slightly more likely than the non-stereotypical, abnormal state $t_{\texttt{abn}}$, which corresponds to (10b).[14] Notice also that I am equating receiver response actions directly with world states (see below and section 3.1). For this context model, we would like our solution concept to uniquely single out the Horn-strategy play:

$$S^* = \left\{ \begin{array}{lcl} t_{\texttt{norm}} & \mapsto & m_{\texttt{shrt}} \\ t_{\texttt{abn}} & \mapsto & m_{\texttt{lng}} \end{array} \right\} \qquad R^* = \left\{ \begin{array}{lcl} m_{\texttt{shrt}} & \mapsto & t_{\texttt{norm}} \\ m_{\texttt{lng}} & \mapsto & t_{\texttt{abn}} \end{array} \right\}.$$

Unraveling M-Implicatures. This is indeed what the IBR model provides, and it follows from the more general result that the basic IBR model accounts for what I will call generalized M-implicatures, of which the M-implicature game in figure 2.3 is the special case $n = 2$. Take a signaling game with n states, n messages and n response actions, $n \geq 2$. Assume furthermore that $t_1, t_2, \ldots, t_{\texttt{n}}$ is strictly decreasing in prior probability and that $m_1, m_2, \ldots, m_{\texttt{n}}$ is strictly increasing in message costs with $\mathrm{C}_{S,R}(m_{\texttt{n}}) < 1$. Finally, assume that actions are *interpretation actions* (see also section 3.1) that can be equated with the set of states $A = T$ because we assume utilities as follows:

$$\mathrm{V}_{S,R}(t,a) = \begin{cases} 1 & \text{if } t = a \\ 0 & \text{otherwise.} \end{cases}$$

For such a game both sequences of the vanilla IBR model reach the same fixed point $\langle S^*, R^* \rangle$ for which $S^*(t_{\texttt{i}}) = m_{\texttt{i}}$ and $R^*(m_{\texttt{i}}) = t_{\texttt{i}}$ for all $1 \leq i \leq n$. An obvious *unravelling argument* establishes this result. I will only sketch it here. Notice, first of all, that $R_0 = R_1$, so that it suffices to show that the R_0-sequence has the fixed point in question. Since R_0 answers all messages

14. Although certain features of it are certainly debatable, this signaling game model is the standard model assumed in game theoretic accounts of Horn's division of pragmatic labor (Parikh 1992, 2001; van Rooij 2004*b*, 2006*b*; Benz and van Rooij 2007; Jäger 2008*c*; de Jaegher 2008; van Rooij 2008) and it is also in line with the standard formalization of the problem in Bidirectional Optimality Theory (see Blutner 1998, 2000, and also chapter 4.1.3).

with t_1, S_1 will never induce her preferred action in any state other than t_1 and so she will always send the cheapest message m_1:

$$S_1 = \left\{ \begin{array}{ccc} t_1 & \mapsto & m_1 \\ t_2 & \mapsto & m_1 \\ \vdots & & \vdots \\ t_n & \mapsto & m_1 \end{array} \right\}$$

This renders all of $m_2, \ldots, m_n$ surprise messages for R_2. The vanilla IBR model does not restrict counterfactual beliefs, and —as we have seen above— therefore collects all actions that are zero-order rationalizable. Since in the present game all actions are rational for some belief in $\Delta(T)$ this yields:

$$R_2 = \left\{ \begin{array}{ccc} m_1 & \mapsto & t_1 \\ m_2 & \mapsto & t_1, t_2, \ldots, t_n \\ \vdots & & \vdots \\ m_n & \mapsto & t_1, t_2, \ldots, t_n \end{array} \right\}$$

Given this receiver behavior it becomes advantageous for S_3 to send the cheapest message other than m_1, i.e., m_2, in all states other than t_1. This will have R_4 respond to m_2 with the most likely state where it is being sent, which is t_2, but it also leaves him surprised by messages $m_3, \ldots, m_n$:

$$S_3 = \left\{ \begin{array}{ccc} t_1 & \mapsto & m_1 \\ t_2 & \mapsto & m_2 \\ t_3 & \mapsto & m_2 \\ \vdots & & \vdots \\ t_n & \mapsto & m_2 \end{array} \right\} \qquad R_4 = \left\{ \begin{array}{ccc} m_1 & \mapsto & t_1 \\ m_2 & \mapsto & t_2 \\ m_3 & \mapsto & t_1, t_2, \ldots, t_n \\ \vdots & & \vdots \\ m_n & \mapsto & t_1, t_2, \ldots, t_n \end{array} \right\}$$

It is clear how this process continues until after $(2 \times n)$ rounds of iteration a fixed point is reached in which every message m_i is associated one-to-one with t_i by sender and receiver behavior. Consequently, the IBR model again solves the problem of equilibrium selection also for Horn's division of pragmatic labor, for it uniquely selects the intuitively desirable equilibrium even for arbitrary generalizations with n states and n messages.

Reflection. This strong prediction, though theoretically neat, also has its opponents. Beaver and Lee (2004) argue in a slightly different context —a discussion of optimality theory for use in linguistic pragmatics (see chapter 4)— that a system that predicts generalized M-implicatures is actually

flawed, because it overgenerates as there is no data in support of this very strong prediction.[15] To this, I have two replies.

Firstly, the IBR model's prediction is *not* confined to the most rational limit behavior. The IBR model presented here predicts idealized pragmatic reasoning, but it also predicts resource-bounded pragmatic reasoning. Obviously, if we don't observe generalized M-implicature play, this is totally in line with the predictions of the IBR model and the results from behavioral game theory reported in section 2.1 of this chapter, that iterated reasoning such as needed for generalized M-implicatures is restricted to a few steps only.

Secondly, and to my mind more importantly, there is yet another performance limitation that may explain the absence of generalized M-implicatures. My preferred interpretation of a signaling game is as a model of the context of utterance, more specifically as the receiver's belief that it is common belief between sender and receiver that the context of utterance is as modelled by the signaling game (see section 3.1). For a generalized M-implicature to occur, it would be required that a hearer may reasonably come to believe that it is common belief that a given form is associated with a long chain of decreasingly complex alternative forms, which are quite possibly fairly unrelated lexical associations. It might therefore also be a natural portion of uncertainty about the context of utterance which prevents generalized M-implicatures from occurring frequently in the wild. Still, I take it to be an *advantage* that the present model lets us derive generalized M-implicatures for *idealized* agents when they are sufficiently certain that *this* is the game that is being played.

Intermediate Summary. To sum up briefly here, the vanilla IBR model explains scalar and M-implicatures by uniquely selecting the empirically attested speaker and interpretation behavior. This much is already a small achievement. Still, there is room for improvement, and therefore the following sections discuss slightly stronger versions of the model. But before coming to that, I would like to briefly reflect on a common assumption shared by all versions of the IBR model, viz., the consistency requirement on the receiver's beliefs.

15. This may then speak in favor of Jäger's (2008) version of the IBR model (see section 2.4) which does account for simple M-implicatures, but not for generalized M-implicatures.

2.2.3 Consistency: Naïve & Sophisticated Updates

Let us have a brief look back at the previous some-all example, in particular at the one non-trivial application of the consistency requirement on the receiver's beliefs. As we have seen, there is only one posterior receiver belief μ_1 that is consistent with a belief in sender strategy S_0. Although the prior probabilities on states were equal, after observing the message $m_{\texttt{some}}$ the posterior $\mu_1(\cdot|m_{\texttt{some}})$ renders the state $t_{\exists\neg\forall}$ twice as likely as the state $t_{\forall}$. This may seem peculiar: why are $\mu_1(t_{\exists\neg\forall}|m_{\texttt{some}})$ and $\mu_1(t_{\forall}|m_{\texttt{some}})$ not equal, given that prior probabilities are equal and given that S_0 sends the message $m_{\texttt{some}}$ in both of these states?

The answer is that consistency requires the receiver to form his posterior beliefs in a *sophisticated* manner, viz., in such a way that a posterior $\mu(\cdot|m)$ that is consistent with some behavioral belief does not only take into account which states send message m, but also which other messages those states that send m might send alternatively. To see this difference, let us define

$$S_k(m) = \{t \in T \mid \exists s \in S_k : \ s(t) = m\}$$

as the set of all states that send m according to sender strategy S_k. We could now say that a receiver of strategic type $k \geq 1$ with belief $\langle \Pr, S_k, \mu_k \rangle$ performs a NAÏVE UPDATE (alternatively: UNSOPHISTICATED UPDATE) if his posterior is derived from Pr and S_k by Bayesian conditionalization on $S_k(m)$:

$$\mu_k(t|m) = \Pr(t|S_{k-1}(m)).$$

In that case we say that the triple $\langle \Pr, S_k, \mu_k \rangle$ is NAÏVELY CONSISTENT. This contrasts with SOPHISTICATED UPDATE, in which the posteriors are required to be consistent *simpliciter*. To give the obvious example: a naïvely updating R_1 in the some-all game would consider both states equally likely after hearing $m_{\texttt{some}}$, while a sophisticated updater would consider $t_{\exists\neg\forall}$ twice as likely as $t_{\forall}$.

The conceptual difference between naïve and sophisticated update is this. A naïve update takes into account which states a message m is sent in, but it does not take into account —as a sophisticated update would do— with which probability m is sent in each state. In other words, a naïve posterior belief $\mu(\cdot|m)$ rests on the (possibly false) assumption that all and only types $t \in S_k(m)$ *always only* send message m. In contrast, a sophisticated posterior belief $\mu(\cdot|m)$ assumes that all and only types $t \in S_k(m)$ *sometimes* send message m, but that these types may also occasionally send different messages with specific probabilities.

The IBR model thus requires sophisticated updating in the receiver's belief formation. Still, the question remains how exactly sophisticated updating works and why it is more adequate than naïve updating.[16] In order to answer this question it pays to introduce the idea of updating on naïve and sophisticated spaces, and to review empirical research on subjects' judgements of conditional probabilities in the laboratory.

BERTRAND'S BOX PROBLEM. Consider a variation of the so-called BERTRAND'S BOX PROBLEM, a well-known puzzle about conditional probabilities (Bar-Hillel and Falk 1982).[17] Suppose that there are three playing cards, the first of which is red on both sides, the second of which is white on both sides, and the last of which is red on one side and white on the other. Now imagine that one card is drawn at random and you only observe one side of that card. For concreteness, let's say that you observe that the visible side of the selected card is red. What is the probability that the other side of that card is also red?

It is tempting to think that the probability is $1/2$. The *naïve argument* for this would go something like this: initially there are three equally likely possibilities because there are three cards all of which are equally likely to be drawn; my observation rules out that the selected card has white on both sides; but that leaves two equally likely possibilities and hence the probability that the other side of the selected card is red is $1/2$.

Alternatively, one could argue that the probability of the other side of the card being red is $2/3$. The *sophisticated argument* for this would then be: since initially I could get to observe each side of each card, there are six equally likely possibilities; when I observe that one side of the chosen card is red, I can eliminate three of those possibilities; but that leaves three possibilities in the race; in one of those three possibilities, the other side of the card is white, while in two of those possibilities the other side of the card is red; hence the probability that the other side of the selected card is red is $2/3$.

Both the naïve and the sophisticated argument start from an assessment of a set of possibilities which are deemed equally likely. Both arguments then

16. The distinction deserves attention also because it will transpire later that some related approaches in formal pragmatics (Blutner 1998; Benz and van Rooij 2007) rely on naïve update where the IBR model subscribes to sophisticated updates (see sections 2.4 and 4.3).

17. Related problems that would show the same point are the "Monty Hall problem", or the equivalent "three prisoners problem." I discuss Bertrand's box problem, because it is easier, its 'normatively correct' solution is more readily acceptable and it relates more directly to updating in the some-all game.

rule out those possibilities that are incompatible with the given observation. The difference between the naïve and the sophisticated argument is that the former conceptualizes the problem on a naïve space, whereas the latter consults a sophisticated space (see Grünwald and Halpern 2003). The *naïve space* only distinguishes three possibilities, viz., which of the three available cards as a whole is observed. The *sophisticated space* distinguishes more fine-grained information, viz., which side of which card is observed.

The normatively correct answer to Bertrand's box problem is $^2/_3$, the answer backed up by the sophisticated argument. If in doubt, the reader could imagine a repeated performance of the problem as an experiment: draw a card at random and look at only one side of it; whenever the card shows red, count the number of times the opposite side turns out to be red and white; whenever the card shows white, put it back and start again. The only cards that would ever enter this counting process are the red-red card (all of the time that it is drawn) and the red-white card (half of the time that it is drawn). Therefore, the count for red will roughly double the count for white in repeated execution of the problem. (If still in doubt, the reader is advised to actually *perform* the count, preferably without gambling on the outcome.)

Priming on Sophisticated Partitions. Being the normatively correct answer does not mean being the answer that many or most people would give in response to such a problem, be they experts or laymen.[18] Indeed, Fox and Levav (2004) found that a majority of subjects seem to judge conditional probabilities in accordance with a naïve updating strategy. But Fox and Levav's study also showed that subjects can be primed into a sophisticated update if the problem statement was presented so as to raise the salience of a sophisticated space.

More in particular, Fox and Levav found empirical support for their hypothesis that subjective conditional probability is assessed by a simple three step partition-edit-count strategy: (i) partition the space of initial possibilities, (ii) remove possibilities inconsistent with the given observation, and then (iii) count the number of remaining possibilities. This algorithm entails that subjects readily adopt flat priors (in the absence of information to the contrary)[19]

18. This is evidenced by a particularly long and heated discussion about the 'true' answer to the Monty Hall problem (see Savant (1994), as well as `http://www.marilynvossavant.com/articles/gameshow.html`).

19. This then also supports my use of flat priors in interpretation games (see section 3.1), as well as the assumption of unbiased beliefs in player's reasoning chains (see section 2.4).

and that they subsequently perform either a naïve or a sophisticated update, depending on how they partitioned the logical space in the first step of the procedure. Fox and Levav's data showed that although under a neutral formulation of a probability problem, such as the Bertrand's box problem or a version of the Monty Hall problem, subjects tend to perform naïve updates, slight rewordings of the problem statement helped trigger sophisticated partitioning by raising the salience of the additional distinctions of the sophisticated space.

Consistency as Update on Sophisticated Spaces. This is relevant also for the IBR model and its assumption about the receiver's belief formation, because sophisticated updating in the IBR model can be conceived of as updating on a sophisticated space that takes the sender strategy into account. A naïve space, on the other hand, does not take the sender strategy into account. The main argument for adoption of sophisticated updating in the IBR model is then that IBR is essentially about reasoning about the opponent's strategy and that it is thus legitimate to assume that the receiver construes a sophisticated space that duly respects the necessary distinctions.

To see what is at stake, take once more the sender's strategy S_0 in the some-all game from above:

$$S_0 = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\texttt{some}} \\ t_{\forall} & \mapsto & m_{\texttt{some}}, m_{\texttt{all}} \end{array} \right\}$$

and consider how either a naïve or a sophisticated R_1 would partition a possibility space and update with the observation $m_{\texttt{some}}$. A naïve receiver would consider two possibilities, equally probable at the outset:

(Poss 1) actual state: $t_{\exists\neg\forall}$

(Poss 2) actual state: $t_{\forall}$

Since both possibilities are compatible with the observation $m_{\texttt{some}}$ under the belief in S_0, nothing is eliminated and the posterior belief after $m_{\texttt{some}}$ equals the prior belief.

Consider, on the other hand, a sophisticated receiver R_1 who takes into account the sender's strategy when individuating possibilities in a sophisticated space. A sophisticated R_1 considers two pure strategies of the sender possible. A sophisticated space would therefore distinguish four contingencies individuated by the product of which state is actual and which pure strategy the sender is playing:

(Poss 1) actual state: $t_{\exists\neg\forall}$
sender plays: $[t_{\exists\neg\forall} \mapsto m_{\texttt{some}};\ t_{\forall} \mapsto m_{\texttt{all}}]$

(Poss 2) actual state: $t_{\exists\neg\forall}$
sender plays: $[t_{\exists\neg\forall} \mapsto m_{\texttt{some}};\ t_{\forall} \mapsto m_{\texttt{some}}]$

(Poss 3) actual state: $t_{\forall}$
sender plays: $[t_{\exists\neg\forall} \mapsto m_{\texttt{some}};\ t_{\forall} \mapsto m_{\texttt{all}}]$

(Poss 4) actual state: $t_{\forall}$
sender plays: $[t_{\exists\neg\forall} \mapsto m_{\texttt{some}};\ t_{\forall} \mapsto m_{\texttt{some}}]$

An observation of message $m_{\texttt{some}}$ is incompatible with only possibility 3, because in this possibility $m_{\texttt{some}}$ would not be sent. But that means that two possibilities with $t_{\exists\neg\forall}$ remain, but only one with $t_{\forall}$. This explains the workings of consistency on a sophisticated space that takes the sender's sending strategy duly into account.

Naïve or Sophisticated Update in Language Interpretation? In conclusion, naïve update appeals because it is easier to calculate. In fact, under a naïve update the whole IBR model becomes much simpler: it is easy to verify that under naïve update we have $R_0 = R_1$ and so the S_0-sequence collapses into the R_0-sequence.[20] Nonetheless there are several reasons to prefer sophisticated update.

Firstly, updating naïvely is often an actual *mistake*. As such it is unlike, for instance, computing only finitely many steps of an IBR sequence. Assuming that reasoners systematically make a particular *mistake* seems like an odd strategy for a model of pragmatic competence. Such an assumption might be defensible if the predictions derived under it would have superior empirical coverage on the to-be-explained data. But this is not so. In fact, my second reason for subscribing to sophisticated updating is that the system's predictions are much better with sophisticated update than those that we would derive with naïvely updating receivers.[21] Thirdly, lastly and most importantly, sophisticated update is also defensible on empirical grounds. As Fox and Levav

20. This is an interesting issue to ponder in the context of the question whether pragmatic interpretation is to start with a naïve sender or with a naïve receiver. Under naïve update this distinction is futile.

21. Anticipating a little, we would, for instance, need extra assumptions, such as non-flat priors implementing minimality of states, in order to account for implicatures of disjunctions, as well as free choice implicatures (see sections 3.2 and 3.3).

(2004) show, it is possible to prime subjects into conceptualizing a sophisticated space if the necessary distinctions are sufficiently salient. In the context of IBR reasoning, it is plausible to assume that reasoners are sufficiently aware of their opponent's strategies and the strategic implications of these. This is, in essence, what the IBR model is basically about. Whence that the assumption that receivers update on a sophisticated space which duly respects the sender's strategy seems legitimate also from an empirical point of view.

2.2.4 Truth Ceteris Paribus & Skewed Priors

The main motivation for the IBR model which I gave in chapter 1 was a proper implementation of semantic meaning into a game theoretic solution concept. So far, conventional meaning has been implemented as a focal point at the beginning of the IBR sequence. This is sufficient for many examples, but still there are also good arguments why the impact of conventional meaning on the IBR reasoning should be strengthened slightly, by what I will call a *truth ceteris paribus* assumption: the idea that the sender will stick to conventional meaning at later stages of the IBR sequence if otherwise indifferent. This section motivates and implements such an extra assumption.

EXAMPLE. Consider again the some-all game in figure 1.3, but assume this time that the prior probabilities are not flat, but rather skewed towards $t_\forall$: let $\Pr(t_\forall) = p > 1/2$. With these priors, the naïve receiver R_0 has the beliefs

μ_0	$t_{\exists\neg\forall}$	$t_\forall$
$m_{\texttt{some}}$	$1-p$	p
$m_{\texttt{all}}$	0	1

to which his best response is to play $t_\forall$ in both states (since $p > 1/2$):

$$R_0 = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_\forall \\ m_{\texttt{all}} & \mapsto & t_\forall \end{array} \right\}.$$

But then S_1 believes that she cannot induce action $t_{\exists\neg\forall}$ by either message in state $t_{\exists\neg\forall}$, and she is thus indifferent between sending a true message $m_{\texttt{some}}$ and an untrue message $m_{\texttt{all}}$. As things stand, S_1 is expected to send either message in either state, and the sequence thereby reaches a pooling fixed point in which $m_{\texttt{some}}$ is interpreted to mean $t_\forall$.

This prediction is not entirely satisfactory. For values of p just slightly bigger that $1/2$, only the R_0-sequence predicts such pooling, while the S_0-sequence predicts the scalar implicature play for all values $p \leq 2/3$ and predicts pooling of this sort only for $p > 2/3$. It would certainly be desirable to have the same fixed point prediction for both sequences, in particular, the stronger scalar implicature prediction also for $1/2 < p \leq 2/3$ in the R_0-sequence, so that the intuitive implicature prediction is not sensitive to slight deviations from flat priors.

Truth Ceteris Paribus. Therefore, it is here that we should strengthen the impact of conventional meaning on pragmatic reasoning slightly. Also intuitively there is something strange about the above pooling outcome. Why would we not trust the conventional meaning of messages in this case, if after all the sender has no positive incentive to deviate from the semantics? It seems that whenever the sender could in principle say something true, when otherwise being indifferent, we may as well expect the sender to stick to the truth. Effectively, I argue, there is an expectation of the interpreter of a secondary preference for *truth ceteris paribus* (TCP) of the speaker.

To implement the TCP assumption as the receiver's expectation about sender behavior in the IBR model we simply have to restrict the strategies of sender type S_{k+1} so that whenever the sender could optimally say something true, she will do so. Formally, the IBR model implements the TCP assumption if we define S_{k+1}, not as $S_{k+1} = \mathrm{BR}(R_k)$, but as:

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

If S_{k+1} is a representation of the unbiased belief of R_{k+2}, this restriction in the inductive definition of the IBR model implements a bias in the receiver's belief formation: the receiver *expects* a true message all else being equal.

TCP as Nominal Costs. The TCP assumption could equivalently be thought of as a *nominal* cost of false signaling. In order to appreciate this, let us first introduce the concept of *nominal message costs*. As we have already seen, we would sometimes like to regard especially the sender's utility function U_S as composed of *response utilities* $\mathrm{V}_S : T \times A \rightarrow \mathbb{R}$ and *message costs* $\mathrm{C}_S : T \times M \rightarrow \mathbb{R}$. One way of combining these is by straightforward subtraction:

$$\mathrm{U}_S(t, m, a) = \mathrm{V}_S(t, a) - \mathrm{C}_S(t, m).$$

More generally, of course, the operation that combines response utilities and message costs need not be subtraction. In general, we could think of the utility function U_S as *some* composite function $F : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ that takes response utilities and message costs and maps these onto a real number giving the preference of the sender:

$$U_S(t, m, a) = F(V_S(t, a), C_S(t, m)).$$

This makes it possible to correctly spell out the idea, for instance, that response utilities are *always* more important to the sender's expected utilities than message costs, i.e., that what matters first and foremost for the sender's expected utility is the response utility, and that only where this is undecided do message costs apply. Formally speaking, this idea says that the speaker's expected utilities are defined in terms of a *lexicographic ordering* that ranks response utilities higher than message costs. A utility function U_S implements nominal message costs if it gives rise to such a lexicographic ordering under expected utility calculation.

More concretely, define the sender's EXPECTED RESPONSE UTILITIES as

$$EV_S(m, t, \rho) = \sum_{a \in A} \rho(a|m) \times V_S(t, m, a).$$

We then say that the sender's utility function U_S implements NOMINAL MESSAGE COSTS if it is a functional combination of response utilities and message costs such that for all t, ρ, m and m' we have

$$\begin{aligned} EU_S(m, t, \rho) > EU_S(m', t, \rho) \quad \text{iff} \quad & \text{(i)} \quad EV_S(m, t, \rho) > EV_S(m', t, \rho) \text{ or} \\ & \text{(ii)} \quad EV_S(m, t, \rho) = EV_S(m', t, \rho) \text{ and} \\ & \qquad C_S(t, m) > C_S(t, m'). \end{aligned}$$

It becomes obvious that we may either think of the TCP assumption as an assumption about belief formation of agents (in particular of the receiver), or alternatively as a uniform nominal message cost for sending false signals. To make sense of the latter approach, we would need to assume that the given utilities of a cheap-talk signaling game are response utilities and that there are nominal message costs

$$C_S(t, m) = \begin{cases} c > 0 & \text{if } t \notin [\![m]\!] \\ 0 & \text{otherwise} \end{cases}$$

that apply to the sender's overall utilities.

Example — Continued. Returning to the example, given belief in R_0 the TCP assumption yields that S_1 is expected to send only the true message $m_{\texttt{some}}$ in state $t_{\exists\neg\forall}$ though otherwise indifferent:

$$S_1 = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\texttt{some}} \\ t_{\forall} & \mapsto & M \end{array} \right\}.$$

Deriving R_2's posteriors from S_1 gives a different result:

$$\begin{aligned} \mu_2(t_\forall | m_{\texttt{some}}) &= \frac{\Pr(t_\forall) \times S_1(m_{\texttt{some}} | t_\forall)}{\sum_{t' \in T} \Pr(t') \times S_1(m_{\texttt{some}} | t')} \\ &= \frac{p/2}{p/2 + 1 - p} \\ &= \frac{p}{2-p}. \end{aligned}$$

We find that now the receiver's best response to message $m_{\texttt{some}}$ will not necessarily be $t_\forall$ anymore. Only if $\frac{p}{2-p} > 1/2$, i.e., if $p > 2/3$, will the receiver interpret $m_{\texttt{some}}$ as $t_\forall$. For values $p < 2/3$ we get

$$R_2 = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_{\exists\neg\forall} \\ m_{\texttt{all}} & \mapsto & t_{\forall} \end{array} \right\}$$

and for $p = 2/3$ we get

$$R_2 = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & T \\ m_{\texttt{all}} & \mapsto & t_{\forall} \end{array} \right\}.$$

Still, the best response of the sender S_3 to either receiver strategy is to play

$$S_3 = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\texttt{some}} \\ t_{\forall} & \mapsto & m_{\texttt{all}} \end{array} \right\}$$

with which the scalar implicature fixed point is reached.

In summary, the presence of the TCP assumption ensures that small deviations from flat priors still result in the intuitively correct scalar implicature prediction for both IBR sequences. Without the TCP assumption, even a minute deviation from a flat prior has the IBR model predict a pooling outcome for the R_0-sequence. With the TCP assumption the predictions of the R_0-sequence exactly match those of the S_0-sequence (which are the same with or without

TCP) for all values of $\Pr(t_\forall)$: in particular, for $\Pr(t_\forall) \leq {}^2/_3$ we get the scalar implicature play; for $\Pr(t_\forall) > {}^2/_3$ we predict the pooling outcome:

$$R^* = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_\forall \\ m_{\texttt{all}} & \mapsto & t_\forall \end{array} \right\}$$

$$S^* = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\texttt{some}} \\ t_\forall & \mapsto & M \end{array} \right\}.$$

In other words, the TCP assumption helps, among other things, to rule out unintuitive pooling behavior from the model's limit prediction that arises from small deviations from flat priors. It it for this reason that I will always adopt this TCP assumption in the basic model.

HEAVILY SKEWED PRIORS. A wrinkle remains. Is it not bad that the IBR model predicts the scalar implicature play only for most constellations of prior probabilities, but not all? Is it not unintuitive that whenever the state $t_\forall$ is more than twice as likely as $t_{\exists\neg\forall}$, that we interpret $m_{\texttt{some}}$ as $t_\forall$? For several reasons I do not think that this is problematic. Let me then briefly enlarge on this here, even at the risk of digression.

First of all, we need to settle the question how prior probabilities in the signaling game context model are to be interpreted. Only then can we answer whether the model's prediction here is intuitive. I argue extensively in section 3.1 that the prior probabilities in a signaling game should *not* be thought of as specifications of the receiver's beliefs about which state is actual *before* he has observed a message. Rather, in a pragmatic context, prior probabilities should be considered condensed and simplified representations of generally accessible meaning associations. From this point of view it is sufficient that the IBR model can deal with slightly and even fairly skewed priors, while predicting pooling for cases of extreme associative biases. I consider these extreme contexts unnatural and I am not worried if a theory makes unintuitive predictions for unnatural, non-occurring parameter settings, as long as there is some sufficient margin around natural parameter settings in which predictions are robust.

Nonetheless, some readers may not like my interpretation of prior probabilities and these readers may find the pooling prediction under heavily skewed priors objectionable. In that case, consolation may be found in a more technical solution.[22] Suppose we introduce an infinitesimal sender uncertainty about the prior probabilities into the model: suppose that the sender

22. The following idea is derived from a proposal by Tikitu de Jager on how to eliminate

considers it very unlikely, but still possible, that the priors are *not* heavily skewed so that at least with infinitesimal ϵ-chance the receiver is expected to respond to unskewed, i.e., flat or nearly flat, priors. This assumption is not unnatural, because there is arguably always some uncertainty about the actual context of utterance. But with this, even if the sender believes that the receiver plays

$$R^* = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_\forall \\ m_{\texttt{all}} & \mapsto & t_\forall \end{array} \right\}$$

it becomes suboptimal to use $m_{\texttt{some}}$ in state $t_\forall$, because by ϵ-chance this message might be interpreted incorrectly as $t_{\exists\neg\forall}$, while the message $m_{\texttt{all}}$ will never be incorrectly interpreted even if the receiver may have different prior probabilities (as long as the receiver can be expected to make no mistakes about semantic meaning). In effect, with a natural arbitrarily small sender uncertainty about the receiver's priors, pooling can be ruled out even for heavily skewed priors.

Summary. Taken together, the IBR model implements semantic meaning most prominently as the focal starting point of best response reasoning. To fine-tune predictions, such as in the light of skewed priors, we would additionally like to require that semantic meaning also impacts the use of messages at later stages of the IBR reasoning. In particular, we would like the sender to send a true message, rather than a false, whenever she is otherwise indifferent. This requirement can be thought of as either an epistemic assumption, a bias in the belief formation of the receiver, or as a nominal cost for untrue signaling. I would like to consider the TCP assumption part of the basic vanilla IBR model because it makes the model's predictions for scalar implicatures robust under varying priors. I argued that heavily skewed priors are unnatural context assumptions to begin with, but may still be dealt with if we also introduce a minimal fragrance of sender uncertainty about the receiver's priors.

2.3 Forward Induction

This section enlarges on a refinement of the basic IBR model, called forward induction assumption, which is a constraint on the receiver's counterfactual

pooling equilibria in the context of finite persuasion games (see Franke et al. to appear). Robert van Rooij repeatedly argued in favor of this idea as a general means of ruling out pooling equilibria.

beliefs. I will first motivate the adoption of such a refinement by appeal to a simple example in section 2.3.1. Section 2.3.2 gives a general introduction to forward induction reasoning and section 2.3.3 shows how to implement this kind of reasoning in the IBR model.

2.3.1 Trouble-Maker "Some But Not All"

Like many other (Neo-)Gricean theories, the vanilla IBR model is vulnerable to a version of the so-called *symmetry problem*.[23] Take an amended some-all signaling game, which is like that in figure 1.3 with $p = 1/2$, but which also includes a message $m_{\texttt{sbna}}$ —short for "some but not all"— with the obvious semantics $[\![m_{\texttt{sbna}}]\!] = \{t_{\exists\neg\forall}\}$. If we assume that all messages are equally costly (or costless), the vanilla model predicts (in the limit) that the message $m_{\texttt{some}}$ is not going to be sent and will also not pragmatically strengthened by a scalar inference. Concretely, the equilibrium play that the model uniquely selects in this case is:

$$S^* = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\texttt{sbna}} \\ t_{\forall} & \mapsto & m_{\texttt{all}} \end{array} \right\} \qquad R^* = \left\{ \begin{array}{lcl} m_{\texttt{sbna}} & \mapsto & t_{\exists\neg\forall} \\ m_{\texttt{some}} & \mapsto & t_{\exists\neg\forall}, t_{\forall} \\ m_{\texttt{all}} & \mapsto & t_{\forall} \end{array} \right\}. \tag{2.1}$$

Here the unspecific message $m_{\texttt{some}}$ is a surprise message and will therefore be responded to with any zero-order rationalizable action. This is clearly not a desirable prediction, although it is sound from a purely analytic point of view.[24]

There are two standard solutions to this problem. Either (i) we could assume that specific forms like $m_{\texttt{sbna}}$ should be excluded from reasonable context models, or (ii) we could argue that whenever such specific forms are included, they incur a small message cost that sets them off from other messages. Using the present game theoretic jargon, both lines of defense are geared towards a proper specification of the context model. Indeed, I will argue for option (i) as a reasonable constraint on models of standard, generic contexts (see section 3.1), but I do not want to rely on option (i) entirely, and

23. I will come back to the symmetry problem in section 3.1.

24. If it's common knowledge between speaker and hearer that those three messages are available to the speaker, all at equal cost and with the assumed semantic meaning, there is indeed *no reason whatsoever* why $m_{\texttt{some}}$ should be enriched to mean either only $t_{\exists\neg\forall}$ or $t_{\forall}$. This is actually an interesting point to notice: unlike, for instance, a diachronic, evolutionary account with a natural small mutation rate, the IBR model does not support *arbitrary* meaning enrichment.

this is where a problem arises for the vanilla model as a solution concept: even if $m_{\texttt{sbna}}$ incurs a small cost —smaller than $1/2$— the vanilla model still predicts the unintuitive strategy profile (2.1).

To my mind, if $m_{\texttt{sbna}}$ incurs a cost, this prediction is analytically dubious for the following reason. Take R^* whose behavior is given in (2.1), and who is surprised by message $m_{\texttt{some}}$. The vanilla model keeps $t_\forall$ as a possible interpretation of the surprise message $m_{\texttt{some}}$, because there is a belief π_R under which this interpretation is rational. The question is, however, is this belief itself rational? For there is a rather compelling reason why the receiver should *not* adopt the (counterfactual) posterior belief after hearing $m_{\texttt{some}}$ that the true state of affairs might *possibly* be $t_\forall$. The receiver should answer to himself the question why the sender has sent a surprise message after all by reasoning that there is just one state, namely $t_{\exists\neg\forall}$, in which the sender could possibly profit from sending the surprise message. In $t_\forall$ there is already a cost-minimal message which successfully communicates this state of affairs, but in $t_{\exists\neg\forall}$ there is not: although there is a message which successfully communicates this state, there is also a cheaper message which *could* communicate this state too, but which at present is not used.

This kind of reasoning is known as *forward induction*. Forward induction reasoning is a restriction on counterfactual beliefs held in response to a surprise message m: roughly speaking, the receiver should not put any positive credence on a state t for which there already is a message m' so that most efficient communicative success is already guaranteed in t (no matter how the surprising m would be interpreted). Put the other way around, the receiver should try to *rationalize* the use of a surprise message if there is a conceivable reason for which the sender would want to deviate from a given play. The following section enlarges on this concept.

2.3.2 Forward Induction and Strong Belief in Rationality

As a general motivating example of forward induction reasoning, consider the hawk-dove game in figure 2.4. In this static game, both row player and column player have a choice between playing hawk h and dove d (as usual, the row player's payoff is given first). Classically, this game represents a situation of conflict about a scarce resource. Both players can choose to behave hawkish so as to selfishly fight for the resource at the expense of physical injury, or play dovish so as to only take as much as the other player is willing to give. Since payoffs here only represent players' interest in obtaining the resource

	h	d
h	-2,-2	2,0
d	0,2	1,1

Figure 2.4: The hawk-dove game

and avoiding physical injury, the absolute best outcome for a player is to play hawk and claim the whole resource for herself while the other player gives in by playing dove. However, if two hawk players meet, they will fight and harm each other, which is the worst outcome for both players. If both players play dove, they peacefully share the resource at stake which is worse than getting all the resource, but better than being injured in a fight.

This game has two (asymmetric) Nash equilibra in pure strategies, $\langle h, d\rangle$ and $\langle d, h\rangle$, with expected payoffs of $\langle 2, 0\rangle$ and $\langle 0, 2\rangle$ respectively. But the game also has a Nash equilibrium in mixed strategies, where both players play hawk with probability $1/3$. The expected payoff for each player if the mixed Nash equilibrium is played is $2/3$. Intuitively speaking, each player would like to coordinate on an equillibrium play that has *her* play hawk, and the opponent play dove. But playing hawk in the absence of any reasonable conjecture about the opponent's behavior and beliefs is risky, because the outcome $\langle h, h\rangle$ is the worst that can happen to both players.

But consider now the following variant of this game, where the row player, which we can identify with the sender S for reasons that will become obvious soon, has the opportunity to inflict some damage *on herself* prior to playing the static hawk-dove game against the column playing receiver R. Assuming that the sender's self-damage equals one util this gives rise to the dynamic game in figure 2.5. Dynamic games like this have been studied extensively in the game theoretic literature (Ben-Porath and Dekel 1992; Shimoji 2002), where the initial self-damaging move is often referred to as MONEY BURNING, to highlight the apparent irrationality of a move that merely harms oneself. Having the chance to burn money or inflict self-damage should make no difference to the analysis of the game, one could argue, because why would a rational agent *ever* choose to hurt herself?

Indeed, *backward induction* (see also section 2.1.2) predicts that the sender should not choose to hurt herself here. Backward induction is an iterative procedure that determines each moving player's optimal choices in each subgame of a dynamic game, starting from the last choice points where players

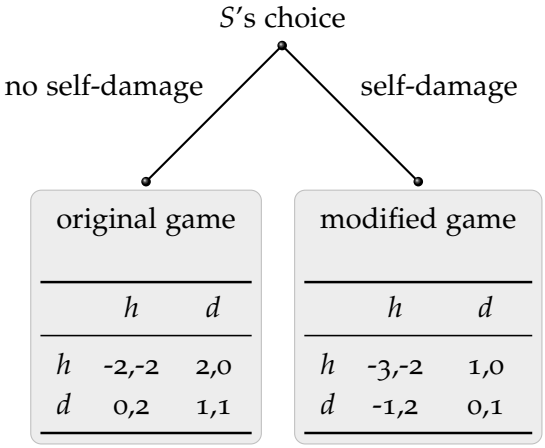


Figure 2.5: The hawk-dove game with initial option of self-damage

move and then propagating optimal choices *backwards* —hence the name— to all earlier choice points. Here's what backward induction does in the above dynamic hawk-dove game. This game has two (strategic) subgames which are strategically equivalent, both of which have the same Nash equilibria: two pure equilibria $\langle h, d \rangle$ and $\langle d, h \rangle$, and one mixed equilibrium in which both players play h with probability $1/3$. The only difference between the two subgames is that S's expected payoff from any equilibrium, pure or mixed, is exactly one util less in the game after she inflicted damage on herself than in the one after she did not. But that means that if the receiver makes his choice in both subgames *independently* of whether the sender chose to hurt herself or not, it would indeed be irrational for the sender to hurt herself. Backward induction predicts exactly that, because backward induction —the name is somewhat unfortunate when we look at things in this way— only looks *forward* into the future moves of the dynamic game and does not take into account the previous game history that led to a particular subgame.

Still, there is ample reason why it may be rational for S to hurt herself after all. S might believe that R would choose to play d if he observes her hurting herself, but would otherwise play h with some positive probability. In that case, it is absolutely rational from S's point of view to inflict damage on herself, because if R indeed plays dove after observing S inflict damage on herself, S actually gains by self-sacrifice after all, because she can play hawk and expect a payoff of 1, where otherwise her expected payoff is strictly smaller than 1.

This is where forward induction reasoning enters. In proper subgames of

a dynamic game, unlike backward induction, forward induction recommends to look *back* —again the unfortunate naming— at the history of the play that led to the given subgame. Forward induction reasoning tries to *rationalize observed behavior*, as much as that is possible. There are many different versions of this idea in the economics literature and it is fair to say that there is no consensus as to what formal notion satisfactorily captures this intuitive reasoning in its entirety. Still, an intuitively very accessible approach to forward induction reasoning is via the following informal *Best Rationalization Principle* of Battigalli (1996):

> Best Rationalization Principle "A player should always believe that her opponents are implementing one of the 'most rational' (or 'least irrational') strategy profiles which are consistent with her information." (Battigalli 1996, p. 180)

This principle captures the essence of forward induction reasoning: we apply forward induction reasoning if we show persistence in the belief that others are rational given any choices that they may have made up to a certain point in time, even after they have failed to choose what seemed to us the most rational option.

Even without further formal specification, it should be clear how the Best Rationalization Principle gives rise to the intuitive verdict in the extended hawk-dove game above: if the receiver adheres to this principle, he should believe that the sender believes that the receiver will play d after the sender has hurt herself, since this is the *only* belief that the receiver could ascribe to the sender which has self-damage come out rational.

This reasoning is intricate, but its bare essentials are, to my mind, intuitive and compelling. I will argue below that the same kind of forward induction reasoning also underlies the hunch that in the some-all signaling game with an additional costly message $m_{\texttt{sbna}}$, the receiver nevertheless comes to interpret message $m_{\texttt{some}}$ correctly as $\{m_{\exists\neg\forall}\}$. The question to be addressed next is how forward induction can be integrated into the IBR model.

2.3.3 Restrictions on Counterfactual Beliefs

The Best Rationalization Principle says that agents ought to persist in their belief in others' rationality as much as possible, even in the face of apparent violations of rationality. This has been implemented in epistemic models of games in order to characterize the notion of rationalizability and iterated dominance (cf. Bernheim 1984; Pearce 1984; Stalnaker 1998; Battigalli

and Siniscalchi 2002). Prior to this, forward induction reasoning was primarily studied as a refinement on equilibrium notions in dynamic games (Kohlberg and Mertens 1986; van Damme 1989). This line of research has spawned a series of more or less complicated restrictions on plausible counterfactual beliefs that may be held in equilibrium. I will focus in the following on the *intuitive criterion* by Cho and Kreps (1987) and I propose that we should implement a similar restriction in the IBR model, which I will call *weak k-dominance*.

THE INTUITIVE CRITERION. Cho and Kreps's *intuitive criterion* is perhaps the most basic restriction of acceptable counterfactual beliefs in equilibria of signaling games that implements pure forward induction reasoning. According to the intuitive criterion, the receiver should not believe that a surprise message was sent by a type that could only do worse by sending that message compared to the given equilibrium (unless this would rule out support of *any* possible state). Consequently, an *intuitive equilibrium* is one which does not rely on beliefs ruled out by the intuitive criterion. More precisely, let $\langle \sigma, \rho, \mu \rangle$ be some perfect Bayesian equilibrium in probabilistic strategies and let $\mathrm{U}^*(t)$ be the sender's expected payoff of that equilibrium in state t. Then define the set of states $T(m)$ that would never want to deviate from the equilibrium outcome by sending a surprise message m, no matter which zero-order rationalizable action (see section 2.2.1) the receiver plays:

$$T(m) = \left\{ t \in T \mid \mathrm{U}^*(t) > \max_{a \in A^*(m)} \mathrm{U}_S(t, m, a) \right\}.$$

We say that the set $T(m)$ is the set of states in which the message m is EQUILIBRIUM DOMINATED. A posterior belief μ satisfies the INTUITIVE CRITERION if for all surprise messages m such that $T(m) \neq T$, it holds that if $t \in T(m)$, then $\mu(t|m) = 0$. This lets us rule out unintuitive equilibria: a perfect Bayesian equilibrium $\langle \sigma, \rho, \mu \rangle$ is an INTUITIVE EQUILIBRIUM if μ satisfies the intuitive criterion.

Although its precise mathematical formulation is rather complex, I believe that the intuitive criterion is a reasonable restriction on counterfactual beliefs, a form of which should also be included in the IBR model. But, unfortunately, the intuitive criterion of Cho and Kreps does *not* help with the some/all/some-but-not-all game that motivated the whole discussion of restrictions on counterfactual beliefs. It turns out that the intuitive criterion is not strong enough for this case. To see this, it suffices to look at the limit prediction of the vanilla IBR model. This is a perfect Bayesian equilibrium, to

which the intuitive criterion should in principle apply. However, the intuitive criterion is vacuous as a restriction on counterfactual belief formation in this case, because the surprise message $m_{\texttt{some}}$, which we would like to have interpreted as $\{t_{\exists\neg\forall}\}$, is equilibrium dominated neither in state $t_{\exists\neg\forall}$, nor in $t_{\forall}$, because the sender does not do *strictly* better in either state by sticking to the equilibrium, so that $T(m_{\texttt{some}}) = \emptyset$. Hence, the intuitive criterion is too weak to place any restrictions on counterfactual beliefs in this case.

Weak k-Dominance. This leaves the problem with the some/all/some-but-not-all game unsolved and us with the wish to possibly consult further, perhaps even stronger refinement notions. But maybe we should not do that. For, as Banks et al. (1994) showed, there seems to be an empirical limit to analytically plausible refinements. Banks et al.'s experimental results suggest that in laboratory experiments subjects conform with the predictions of perfect Bayesian, intuitive equilibrium and possibly divinity (see below) —reasoning themselves towards the more refined equilibrium, if a less refined one also exists— but not necessarily much further.[25]

That is why I will opt for adopting a slightly modified version of the intuitive criterion into the IBR model, which I will call *weak k-dominance*. Since weak k-dominance adds forward induction reasoning to the IBR model, I will refer more generally to this restriction on counterfactual belief formation as the FORWARD INDUCTION ASSUMPTION, or FI assumption for short, which is being added to the vanilla IBR model.

Now, first of all, it should be pointed out that it is not a problem that the intuitive criterion is strictly speaking a refinement of equilibrium: forward induction reasoning does not strictly require equilibrium, but is sound for *any* triple $\langle \sigma, \rho, \mu \rangle$ such that ρ is rational given μ, and μ is consistent with σ; it is not necessary for the intuitive argument that σ is a best response to ρ. So we could, in principle, take over the intuitive criterion, or an amended version of it, as a restriction on counterfactual belief formation of the receiver *at each step* in the IBR model. Still, there is a slight difference in whether forward induction reasoning applies to equilibrium or a sequence of iterated best responses and it therefore pays to have a detailed look at what it means

25. Banks et al. (1994) also tested universal divinity (Banks and Sobel 1987), a notion called never-a-weak-best-response and stability (Kohlberg and Mertens 1986). The exact conclusion that Banks et al. draw from their data is more carefully hedged than presented here. (Of course.) The interested reader is referred to their paper as well as the overview in chapter 8 of Camerer's book (Camerer 2003).

exactly to incorporate forward induction reasoning into the IBR model.

The IBR model defines a series of non-trivial receiver beliefs π_{R_k} for all $k \geq 1$ which includes a belief in the sender's strategy S_{k-1}. This component determines which messages the receiver will consider surprising. It is easily seen that there are no surprise messages given belief in S_0: counterfactual beliefs arise no earlier than for R_2. So for $k \geq 2$, R_k believes that his opponent plays according to S_{k-1} as a rational response to the belief that he plays R_{k-2}. That means that R_k believes that S_{k-1} has expected utility $\mathrm{EU}_s(\cdot, \cdot, R_{k-2})$ and that she sends some message that would maximize this expected utility in the state that she knows to be actual. But that in turn means that if R_k is surprised by a message, something about his belief must be wrong: either S_{k-1} does not believe in R_{k-2} or she is not making a rational decision. Enter the Best Rationalization Principle, according to which R_k should preferably *not* revise his belief in S's rationality, but rather revise his belief about S's beliefs. So, according to the principle, R_k should adopt a belief that has the surprise message come out rational after all.

This could in principle be done by many belief ascriptions, but I would like to make and defend what is perhaps a highly contestable simplification, namely the assumption that the rationalization of surprise messages is guided by beliefs about types in the IBR sequence. In other words, whenever R_k needs to rationalize the use of a surprise message, the beliefs that he may adopt about his opponent will be restricted to sender types that occur in the IBR model. That means in particular that if R_k is surprised by S's choice of message given his belief in S_{k-1}, I will assume that R can do either of two things:

1. R_k can rationalize *down* the IBR sequence, so to speak, thinking that he has *overestimated* his opponent who is of a type *lower* than $k-1$; or

2. R_k can rationalize *up* the IBR sequence, thinking that he has *underestimated* his opponent who is of a type *higher* than $k-1$.

Notice that it is always possible for R_k to come to believe that he has overestimated his opponent: R_k can always make sense of surprise messages by coming to believe that S is of strategic level zero. However, this is *not* necessarily really a rationalization of the sender's behavior, because level-zero senders are not necessarily rational. If, on the other hand, the receiver rationalizes a surprise message *up* the IBR sequence, this seems to contradict the assumption that agents are resource bounded, since after all the assumption

was that finite-level types might have reasoning limitations that force them to adopt the belief in a finite-level opponent.

What this suggests is that true forward induction reasoning, if it takes place *within* the IBR model, which —I repeat for clarity— is a simplifying assumption, involves the receiver rationalizing *up* the IBR model. But, in order to conserve the spirit of resource boundedness, upwards rationalization should go only slightly further up the IBR sequence, because of the difficultly involved in this reasoning. More concretely, the forward induction assumption which I propose to include into the model is this: if R_k uses forward induction to rationalize a surprise message m, he will tentatively adopt the belief that S is of type S_{k+1} and reason based on his own interpretation of non-surprise messages which incentives S_{k+1} could have had to send a surprise message. This is then essentially forward induction reasoning folded into the IBR model.

Consequently, I suggest that R_k's counterfactual beliefs are subject to the following formal requirement. Say that a message m which surprises R_k is WEAKLY k-DOMINATED IN STATE t iff there is a message m' which does not surprise him such that there is no zero-order rationalizable receiver action $a \in A^*(m)$ for which $\mathrm{U}_S(t, m, a) > \mathrm{EU}_S(m', t, R_k)$. The set of states in which surprise message m is weakly k-dominated is thus:

$$T_k(m) = \left\{ t \in T \mid \mathrm{U}_k^*(t) \geq \max_{a \in A^*(m)} \mathrm{U}_S(t, m, a) \right\}$$

where $\mathrm{U}_k^*(t)$ is the maximal expected payoff of S_{k+1}^t if she sends a message that does not surprise R_k. By the FI assumption, R_k should then *not* put positive credence on states t after a surprise message m if m is weakly k-dominated, if that is possible: if m is a surprise message for R_k such that $T_k(m) \neq T$, then R_k's posterior beliefs satisfy weak k-dominance if $\mu(t|m) = 0$ for all $t \in T_k(m)$.

EXAMPLE: SOME BUT NOT ALL. The gist of weak k-dominance becomes clear under a simple example. We should check that weak k-dominance solves our initial problem with the some/all/some-but-not-all game. Here, message $m_{\texttt{some}}$ is a surprise message to R_2. However, we can check that $m_{\texttt{some}}$ is weakly 2-dominated in $t_\forall$ but not in $t_{\exists\neg\forall}$. Message $m_{\texttt{some}}$ is weakly 2-dominated in $t_\forall$, because there is a non-surprise message $m_{\texttt{all}}$ which gives the sender her maximal payoff under interpretation R_2. This is the minute difference between *weak* k-dominance and the intuitive criterion: since the sender cannot hope to

do better in state $t_\forall$ the receiver excludes this state, reasoning that there is no point risking misunderstanding when nothing can be gained. But $m_{\texttt{some}}$ is *not* weakly 2-dominated in $t_{\exists\neg\forall}$ because there is no non-surprise message which yields a higher payoff for the sender than when she sends $m_{\texttt{some}}$ and this is interpreted as $t_{\exists\neg\forall}$. Hence, R_2 will compute the scalar implicature for the surprise message $m_{\texttt{some}}$. In a next step, S_3 will then of course use only messages $m_{\texttt{some}}$ and $m_{\texttt{all}}$, and the IBR sequence reaches a fixed point.

EXAMPLE: M-IMPLICATURES. A similar argument shows that the model with forward induction requires fewer iteration steps than the vanilla model to calculate the M-implicature in a setting with two states and two forms. Take the basic example in figure 2.3. It is enough to look at the R_0-sequence, because $R_0 = R_1$. Since S_1 will use the cheap message $m_{\texttt{shrt}}$ in both states, the costly message $m_{\texttt{lng}}$ is a surprise for R_2. Whereas in the vanilla model we then had $R_2(m_{\texttt{lng}}) = T$, we can now use forward induction. Indeed, the message $m_{\texttt{lng}}$ is weakly 2-dominated in $t_{\texttt{norm}}$ because there is a message $m_{\texttt{shrt}}$ which is better for the sender in $t_{\texttt{norm}}$ under interpretation R_2 no matter how $m_{\texttt{lng}}$ would be interpreted. Consequently, the model with forward induction yields

$$R_2 = R^* = \left\{ \begin{array}{lcl} m_{\texttt{shrt}} & \mapsto & t_{\texttt{norm}} \\ m_{\texttt{lng}} & \mapsto & t_{\texttt{abn}} \end{array} \right\}.$$

This is to say that forward induction reduces the iteration steps necessary for the computation of M-implicatures in a setting with two states and two forms from 4 to 2 — a welcome improvement also.

In general, forward induction reduces the number of steps necessary to establish M-implicatures in settings with n states and n forms from $(2 \times n)$ steps without forward induction to $(2 \times (n-1))$ steps with weak k-dominance. The calculation of the generalized M-implicature game under the basic model (see section 2.2.2) changes only slightly. Under the basic model, a level-$(2i)$ receiver, $0 < i < n$, interprets as follows:

$$R_{2i} = \left\{ \begin{array}{lcl} m_1 & \mapsto & t_1 \\ m_2 & \mapsto & t_2 \\ \vdots & & \vdots \\ m_{\texttt{i}} & \mapsto & t_{\texttt{i}} \\ m_{\texttt{i}+1} & \mapsto & t_1, t_2, \ldots, t_{\texttt{n}} \\ \vdots & & \vdots \\ m_{\texttt{n}} & \mapsto & t_1, t_2, \ldots, t_{\texttt{n}} \end{array} \right\}.$$

If we allow weak $(2i)$-dominance to affect the counterfactual beliefs of a level-$(2i)$ receiver, we rather get:

$$R'_{2i} = \left\{ \begin{array}{lcl} m_1 & \mapsto & t_1 \\ m_2 & \mapsto & t_2 \\ \vdots & & \vdots \\ m_{\mathtt{i}} & \mapsto & t_{\mathtt{i}} \\ m_{\mathtt{i}+1} & \mapsto & t_{\mathtt{i}+1}, t_{\mathtt{i}+2}, \ldots, t_{\mathtt{n}} \\ \vdots & & \vdots \\ m_{\mathtt{n}} & \mapsto & t_{\mathtt{i}+1}, t_{\mathtt{i}+2}, \ldots, t_{\mathtt{n}} \end{array} \right\}.$$

Obviously, weak k-dominance only excludes states that have, in a manner of speaking, already been associated with a message at that point of reasoning. This shows that weak k-dominance does not take into account the prior probability of the states that are to be compared.

Prior Probabilities and Divinity. If we wanted the receiver to further take the prior probabilities of states into account when forming counterfactual posteriors, we could of course do that.[26] Another prominent refinement concept that is stronger than the intuitive criterion —that in fact subsumes it— and that does take prior probabilities of states into account is divinity by Banks and Sobel (1987). Roughly speaking, while the intuitive criterion only excludes certain states from the counterfactual beliefs of the receiver, divinity additionally specifies which states that are not ruled out entirely from posterior beliefs should be considered more likely to have sent the surprise message in question than others. Divinity as an equilibrium refinement is technically rather involved (see also Sobel to appear, for an accessible reformulation and comparison). For our present purposes, we can sidestep the technical details and just layer on top of weak k-dominance the additional requirement that prior probabilities be taken into account. More concretely, let's say that a level-k receiver's posterior beliefs μ_k satisfy divine k-dominance if it satisfies weak k-dominance and furthermore satisfies the constraint:

$$\mu_k(t|m) \leq \mu_k(t'|m) \text{ iff } \Pr(t) \leq \Pr(t')$$

for all surprise messages m and states t, t' for which $\mu(\cdot|m) \neq 0$ by weak k-dominance.

26. This extension is not needed for any application in the remainder of this thesis. I mention this for completeness only.

Summary & Reflection. What exactly is the difference between the vanilla IBR model and a more advanced model with an additional forward induction assumption? It will become clear in section 2.4.3 that already the basic IBR model includes a particular forward induction rationale. Weak k-dominance, however, is even stronger, and is specifically needed in pragmatic applications to rationalize costly messages, and, in a manner of speaking, to abbreviate IBR reasoning. This surfaced in the two previous examples.

Beyond these technical arguments for an FI assumption, there are also conceptual reasons why explicit integration of forward induction reasoning is sensible for a theory of pragmatic interpretation. We could think of forward induction in signaling games as a particularly technical implementation of reasoning towards speaker relevance (see Franke et al. to appear). At heart, forward induction is reasoning based on a persistent belief that observed behavior is rational and purposeful. On an abstract level, the parallel to a hermeneutic presumption of rationality that fundamentally underlies natural language understanding is evident. To establish the meaning of an apparent *non-sequitur* as, for instance, in the classical example (18) of Grice (Grice 1989, p. 32), the hearer needs to rationalize the speaker's linguistic behavior, especially where it deviates from expectation; the question to be asked and answered is: "which beliefs of the speaker justify best that she said *that*?"

(18) A: I am out of petrol.
B: There is a garage round the corner.

Thus conceived, a forward induction assumption implements the receiver's attempt to make sense of utterances from the speaker's perspective, asking under which circumstances, in which frame of mind the speaker could have benefited from acting as she did.[27] Even more abstractly speaking FI reasoning is a particular instance of our general intellectual faculty of 'making sense' of the world around us as purposeful:

> "Winnie-the-Pooh sat down at the foot of the tree, put his head between his paws and began to think. First of all, he said to himself: 'That buzzing-noise means something. You don't get a buzzing-noise like that, just buzzing and buzzing, without its meaning something. If there's a buzzing-noise, somebody's making a buzzing-noise, and the only reason for making a buzzing-noise that *I* know of is because you're a bee.' Then

27. Chapter 5 uses such relevance-based forward induction reasoning to explain some aspects of the pragmatic interpretation of conditionals.

> he thought another long time, and said: 'And the only reason for being a bee that I know of is making honey.' And then he got up, and said: 'And the only reason for making honey is so as *I* can eat it.' So he began to climb the tree." (Milne 1991, p. 18)

2.4 Overview and Comparison

It is high time to take stock, to summarize the various versions of the IBR model that this chapter introduced and to compare these to other related models proposed in game theory and game theoretic pragmatics.

2.4.1 Versions of the IBR Model

IBR Scaffolding. A general scheme for a basic IBR model is given in figure 2.6. An IBR model gives an inductive definition of sender and receiver types. As for the base case, we would like to restrict the set of strategies that we start out with in some suitable way, choosing subsets $S_0 \subseteq S$ and $R_0 \subseteq R$. In the inductive step, we compute best responses to some belief that the opponent is of a lower level of strategic sophistication. Here some variation is possible in how this belief is formed, as we will see shortly. All the types defined by an IBR sequence are part of the solution of the model in a broad sense, because this is what the model can make sense of, possibly under a belief in the opponent's bounded rationality. The set of strategies that are infinitely repeated in such a sequence are the model's limit prediction. These strategies are consistent with common belief in rationality and could be regarded as the model's solution in a narrow sense.

Different versions of IBR models result from different assumptions about the inductive base and, more crucially even, the inductive step. Here, a multiplicity of additional assumptions about agents' belief formation can be fed into the model. These different assumptions not only give rise to possibly different predictions, but may also differ conceptually in the sense that they implement weaker or stronger reasoning rationales.

IBR Variety. The vanilla IBR model that I have proposed in this chapter adds several specific assumptions to our basic scaffolding. To begin with, we have identified semantic meaning as a focal strategy at the outset of IBR reasoning. So, the initial restriction on sender and receiver behavior in the inductive base

$$
\begin{array}{lll}
\text{Base:} & \mathsf{S}_0 \subseteq \mathsf{S} & \\
& \mathsf{R}_0 \subseteq \mathsf{R} & \\
\text{Step:} & \mathsf{S}_{k+1} = \{s \in \mathsf{S} \quad | \quad \exists\, \rho \in \Pi_S \ : & \\
& \qquad \text{(S1)} \quad \rho \text{ is a belief based in some fashion on } \mathsf{R}_k & \\
& \qquad \text{(S2)} \quad s \in \mathrm{BR}(\rho)\ \} & \\
& \mathsf{R}_{k+1} = \{r \in \mathsf{R} \quad | \quad \exists\, \pi_R = \langle \Pr, \sigma, \mu \rangle \in \Pi_R \ : & \\
& \qquad \text{(R1)} \quad \sigma \text{ is a belief based in some fashion on } \mathsf{S}_k & \\
& \qquad \text{(R2)} \quad \pi_R \text{ is consistent} & \\
& \qquad \text{(R3)} \quad r \in \mathrm{BR}(\mu)\ \} & \\
\text{Limit:} & \mathsf{S}^* = \left\{s \in \mathsf{S} \mid \forall i\, \exists j > i \ :\ s \in \mathsf{S}_j\right\} & \\
& \mathsf{R}^* = \left\{r \in \mathsf{R} \mid \forall i\, \exists j > i \ :\ r \in \mathsf{R}_j\right\} & \\
& \mathrm{IBR} = \langle \mathsf{S}^*, \mathsf{R}^* \rangle &
\end{array}
$$

Figure 2.6: Basic scaffolding of an IBR model

became the set of all truthful sender strategies

$$\mathsf{S}_0 = \{s \in \mathsf{S} \mid \forall t \ :\ t \in [\![s(t)]\!]\}$$

and the set of all best responses to a literal interpretation of messages

$$
\begin{aligned}
\mathsf{R}_0 &= \mathrm{BR}(\mu_0) \\
\mu_0(m) &= \Pr(\cdot \mid [\![m]\!]).
\end{aligned}
$$

Secondly, the vanilla IBR model makes two specific assumptions about agents' belief formation in the inductive step, namely that (i) agents of level k believe to face an opponent of *exactly* level $(k-1)$, and that (ii) agents of level k think that any possible level-$(k-1)$ behavior is equally likely. With these assumptions of (i) *myopic overconfidence* and (ii) *unbiased beliefs*, as I will call them, the vanilla IBR model therefore has the following induction step:[28]

$$
\begin{array}{lll}
\mathsf{S}_{k+1} = \{s \in \mathsf{S} \quad | \quad \exists\, \rho \in \Pi_S \ : & \\
\qquad \text{(V-S1)} \quad \rho = \mathsf{R}_k & \\
\qquad \text{(S2)} \quad\ \ s \in \mathrm{BR}(\rho)\ \} & \\
\mathsf{R}_{k+1} = \{r \in \mathsf{R} \quad | \quad \exists\, \pi_R = \langle \Pr, \sigma, \mu \rangle \in \Pi_R \ : & \\
\qquad \text{(V-R1)} \quad \sigma = \mathsf{S}_k & \\
\qquad \text{(R2)} \quad\ \ \pi_R \text{ is consistent} & \\
\qquad \text{(R3)} \quad\ \ r \in \mathrm{BR}(\mu)\ \} &
\end{array}
$$

28. As for notation, remember that unbiased beliefs are flat probability distributions over a given set X of opponent strategies, so that it is feasible to write $\rho = \mathsf{R}_k$, for instance, since R_n completely defines the unbiased probabilistic strategy ρ whose support is exactly R_n.

In addition to that, two refinements were motivated in this chapter: (i) the TCP assumption that has the sender stick to conventional meaning unless she strictly profits from deviation (section 2.2.4), and (ii) the FI assumption that refines the receiver's belief revision policies (section 2.3). These extra assumptions can be implemented in further refinements of the inductive step as follows:

$$
\begin{array}{lll}
\mathsf{S}_{k+1} = \{s \in \mathsf{S} & | & \exists\, \rho \in \Pi_S \ : \\
& & \quad \text{(V-S1)} \quad \rho = \mathsf{R}_k \\
& & \quad \text{(S2)} \quad s \in \mathrm{BR}(\rho) \\
& & \quad \text{(TCP)} \quad \forall t\ (\exists s' \in \mathrm{BR}(\mathsf{R}_k)\ t \in [\![s'(t)]\!]) \rightarrow t \in [\![s(t)]\!]\ \} \\
\mathsf{R}_{k+1} = \{r \in \mathsf{R} & | & \exists\, \pi_R = \langle \mathrm{Pr}, \sigma, \mu \rangle \in \Pi_R \ : \\
& & \quad \text{(V-R1)} \quad \sigma = \mathsf{S}_k \\
& & \quad \text{(R2)} \quad \pi_R \text{ is consistent} \\
& & \quad \text{(R3)} \quad r \in \mathrm{BR}(\mu) \\
& & \quad \text{(FI)} \quad \pi_R \text{ satisfies weak } (k+1)\text{-dominance}\ \}
\end{array}
$$

In the remainder of this thesis, especially when applying the model to pragmatics, I will always assume that TCP is in place and that, as a default, FI is absent. The IBR model with TCP and without FI is what I call the basic model, or vanilla model. This way, the workings of FI can be assessed separately.

2.4.2 Related Models

Other noteworthy variations on the basic IBR schema result if we adopt different assumptions about belief formation about opponent behavior, and possibly also different notions of best response calculation. Let me just briefly mention few salient alternatives.

Optimal Assertions. The IBR model proposed here originally arose from criticism of the *optimal assertions* framework spelled out by Benz (2006) and Benz and van Rooij (2007). In the optimal assertions framework, implicatures are calculated based on the idea that a given assertion was optimal, i.e., rational under a literal interpretation. This idea is reminiscent of a limited R_0-S_1-R_2-sequence of the IBR model. There are, however, formal differences in set-up. The biggest difference is that the interpretation of the receiver R_2 in the IBR model implements a sophisticated update with the full sender strategy S_1 (see section 2.2.3). The optimal assertions framework, at least in the formulation given by Benz and van Rooij (2007), instead derives the following

pragmatic interpretation operator (for a signaling game with honest senders and interpretation actions that corresponds one-to-one with states):

$$\begin{aligned} \text{Prag}(m) &= \{t \in [\![m]\!] \mid m \text{ is optimal in } t\} \\ &= \{t \in [\![m]\!] \mid m \in \text{BR}(R_0)(t)\} \end{aligned}$$

This interpretation operator differs from R_2 (on otherwise the same restricted class of games) foremost in that R_2 takes into account the prior probability of states and also the frequency with which S_1 sends messages in different states. Clearly, this interpretation operator then implements a naïve update (see also Franke 2008*a*, for comparison of IBR, optimal assertions and BIOT).

MYOPIA VS. DISTRIBUTED UNSOPHISTICATION. The assumption of myopic overconfidence that each agent believes to be *exactly* one level more sophisticated than their opponent makes the model perspicuous and tractable, but it is also somewhat unrealistic. Why would an agent that can perform, say, up to three level of ToM reasoning necessarily believe that her opponent is performing *exactly* two? Why should she not be uncertain whether her opponent could do none, one or two? In this case the agent would still be overconfident, but no longer myopic. We could then also assume that S_{k+1}, for instance, adopts some belief $\rho \in \Delta(\bigcup_{i \leq k} \mathsf{R}_i)$.

The cognitive hierarchy model of Camerer et al. (2004) does exactly this. Camerer et al. assume that agents of level k are overconfident in believing that their opponent is *at most* of level $(k-1)$. In order to restrict such overconfident belief in reasonable ways, each agent has a conjecture about the distribution of strategic types in the population. Camerer et al. specifically assume that the distribution of strategic types in a population is a Poisson distribution, and that every level-k player derives his population estimate by conditionalizing the population distribution to types strictly lower than k.

Although a cognitive hierarchy model with population estimates is clearly more realistic and provides a good fit of empirical data for laboratory experiments on a wealth of strategic games (cf. Camerer 2003; Costa-Gomes et al. 2009), it is also much more mathematically involved than other strategic type models that implement myopic overconfidence (e.g Stahl and Wilson 1995; Nagel 1995; Crawford 2003). Still, for our present purposes, the simple model that sticks to myopically overconfident players seems good enough. This will be demonstrated by the model's predictions in many linguistically relevant test cases. Whether ultimately a more complex model of interlocutor's belief formation is necessary is an empirical issue. In order to address this issue we

would have to take in particular the kinds of signaling games studied here into the behavioral economist's laboratory — a line of experimentation that has, to the best of my knowledge, not been executed so far. In other words, in the absence of compelling (empirical) evidence against myopia, I will adopt it for the sake of a simpler model.

Non-Exclusive Beliefs. We could scrutinize not only myopia, but also unbiased beliefs, the other assumption that underlies the IBR model's belief formation process. Indeed, Gerhard Jäger has independently suggested an IBR model in which the assumption of unbiased beliefs is relaxed (Jäger 2008*c*; Jäger and Ebert 2009). Jäger's model allows arbitrary *non-exclusive beliefs*, i.e., arbitrary conjectures about opponent play as long as these do not exclude a possibility altogether.[29] Formally, let $\Delta^+(X)$ be the set of all probability distributions on some set X with full support on X:

$$\Delta^+(X) = \{\delta \in \Delta(X) \mid \forall x \in X \ : \ \delta(x) \neq 0\}\,.$$

With this, the assumption of non-exclusive biases yields a definition of the induction step as follows:

$$\begin{array}{lll}
\mathsf{S}_{k+1} = \{s \in \mathsf{S} & \mid & \exists\, \rho \in \Pi_S \ : \\
& & \text{(J-S1)} \quad \rho \in \Delta^+(\mathsf{R}_k) \\
& & \text{(S2)} \quad\ \ s \in \mathrm{BR}(\rho)\ \} \\
\mathsf{R}_{k+1} = \{r \in \mathsf{R} & \mid & \exists\, \pi_R = \langle \Pr, \sigma, \mu\rangle \in \Pi_R \ : \\
& & \text{(J-R1)} \quad \sigma \in \Delta^+(\mathsf{S}_k) \\
& & \text{(R2)} \quad\ \ \pi_R \text{ is consistent} \\
& & \text{(R3)} \quad\ \ r \in \mathrm{BR}(\mu)\ \}.
\end{array}$$

What is the difference between unbiased and non-exclusive beliefs? First of all, we might say that the former are, in a sense, a simpler modelling choice than the latter, because non-exclusive beliefs involve a further quantification

29. There are other differences too. For one, Jäger's model only consults the R_0-sequence. This way we cannot account for unsophisticated sender behavior, for instance, in language acquisition (see section 4.4). For another, the Jäger model is defined for games that are more complex than signaling games. Jäger includes "contexts" as a parameter for sender uncertainty too. His implementation of sender uncertainty differs from the one presented in section 3.2 in relevant ways: Jäger requires best responses to be rational for every possible context. This is actually a non-standard treatment of players' uncertainty under Bayesian rationality, but one that is certainly worthwhile exploring for application to pragmatics, in particular matters of context-sensitivity and the like.

	a_1	a_2	a_3	m_1	m_2	m_3
t_1	1,1	0,0	0,0	$\surd$	$-$	$\surd$
t_2	0,0	1,1	0,0	$-$	$\surd$	$\surd$
t_3	0,0	0,0	1,1	$\surd$	$\surd$	$\surd$

Figure 2.7: Game in which non-exclusive biases are too weak

over beliefs, i.e., probability distributions over a set X, of which there are uncountably many (in non-trivial cases $|X| > 1$). On the other hand, if we look at things from a purely formal perspective, to require unbiased beliefs for a strategy to be rationalized at some iteration step is *strictly stronger* than to require non-exclusive beliefs: obviously every unbiased belief is non-exclusive, but not vice versa. Hence we would expect to find models for which Jäger's model includes more strategies than my IBR model with unbiased beliefs.[30]

Indeed, non-exclusive biases can sometimes turn out too inclusive. Although Jäger's model deals with scalar implicatures, as well as M-implicatures with two states and two messages, it does not account for generalized M-implicatures (see Jäger 2008*c*). Another example where predictions differ is the game in figure 2.7. This game has two specific messages, m_1 and m_2, which are true only in two of the three states, and one universally true message m_3. Since this game is essentially a coordination game for proper interpretation of the state, it is fairly intuitive to except a fully revealing communication outcome with strategy profile:

$$S^* = \left\{ \begin{array}{lcl} t_1 & \mapsto & m_1 \\ t_2 & \mapsto & m_2 \\ t_3 & \mapsto & m_3 \end{array} \right\} \qquad R^* = \left\{ \begin{array}{lcl} m_1 & \mapsto & a_1 \\ m_2 & \mapsto & a_2 \\ m_3 & \mapsto & a_3 \end{array} \right\}.$$

This is indeed exactly what the IBR model predicts under unbiased beliefs. With the more liberal non-exclusive biases, however, we do not predict revealing communication in this simple example. Suffice it to take the R_0-sequence.

30. Notice that it is *not* the case, though, that the limit solution selected by Jäger's model is necessarily a superset of the limit prediction of the IBR model with unbiased beliefs. Each IBR model defines a different sequence through the strategy space, so to speak, and these sequences may diverge substantially for some games.

Starting with a literal interpretation:

$$R_0 = \left\{ \begin{array}{lcl} m_1 & \mapsto & a_1, a_3 \\ m_2 & \mapsto & a_2, a_3 \\ m_3 & \mapsto & a_1, a_2, a_3 \end{array} \right\}$$

we need to ask which non-exclusive beliefs in R_0 may rationalize a sender strategy. It turns out that in Jäger's system every truthful sender strategy is rational for some non-exclusive belief in R_0:

$$S_1 = \left\{ \begin{array}{lcl} t_1 & \mapsto & m_1, m_3 \\ t_2 & \mapsto & m_2, m_3 \\ t_3 & \mapsto & m_1, m_2, m_3 \end{array} \right\}.$$

For instance, a sender of type t_1 may believe that R_0 plays a_1 after m_1 with high probability, and that no other message induces R_0 to play a_1 with noteworthy probability. But she might as well believe that m_3 is the message which induces a_1 with highest probability. Both messages are thus rational in t_1 under some non-exclusive belief. Similar reasoning then leads to $R_2 = R_0$, for which the sequence enters a fixed point. The IBR model with non-exclusive beliefs predicts no pragmatic enrichment and no revealing communication here.

We would not have be worried too much that non-exclusive beliefs are too weak to account for a seemingly arbitrary example like the game in figure 2.7, but it will turn out that this game actually captures part of the essential structure of examples that are central to the concern of pragmatics. As will be apparent later in chapter 3, the game in figure 2.7 reoccurs, so to speak, embedded in context models for the interpretation of disjunctions. It is here, then, that we expect non-exclusive biases to be too weak to yield intuitive predictions in pragmatically relevant cases too.

Best Responses in Learning and Evolution. Best response models that are at least superficially similar to the present variety have also been entertained as models of belief learning in repeated play, as well as in models of (social) evolution of behavioral patterns (see Fudenberg and Levine 1998).

When playing a game repeatedly, players can form, maintain and revise beliefs about opponent behavior based on past observations. Belief-learning models can be quite complex, for instance, by computing a weighted frequency of occurrence of opponent strategies, with weights favoring either recent or initial observations. A very simple version of such belief-learning

models of repeated play is the *Cournot best-response dynamics* in which players believe that their opponents play exactly the same move as in the previous round. By then playing a best response to this belief a sequence of best responses ensues that is superficially similar to the IBR sequence, but not quite the same. Firstly, unlike the IBR model, the Cournot best-response dynamics starts with a random strategy. Secondly, in signaling games —dynamic games with *incomplete* information— it is actually not possible to observe the opponent's whole strategy in one or a few rounds of actual playing: the receiver, for instance, can only observe the message that was sent at a given occasion, but not the whole strategy, a function from states to messages. So, if the IBR model computes best responses to sets of possible strategies as it does, it is not very plausible to interpret this as a sequence of best responding to the opponent's previously observed move. It is at best a sequence of *fictitious play*, a soliloquy of agents reasoning to themselves before playing the game. Thirdly, there is another conceptual reason why the IBR model is *not* a model of naïve learning and adaptation, but rather a model of strategic thinking. The IBR model does not necessarily assume that players develop into more sophisticated types by (a few rounds of) repeated play. Rather the IBR model implements what is possibly a fundamental ToM reasoning capability that needs to develop in young children, and may even fail fully competent adults occasionally (see more in chapter 4.4).

Yet another diachronic variation of IBR is as a model of social evolution under so-called *best-response dynamics* in which at every moment a small fraction of the population plays a best response to the present population distribution (Gilboa and Matsui 1991; Matsui 1992).[31] We could think of IBR models, as entertained here, as special cases of such best-response dynamics in discrete time, in which not just a fraction, but the whole population plays best responses to the present population behavior. It is an interesting topic for future research to link an IBR model of the current variety to diachrony and to apply it to language change and evolution.

2.4.3 IBR vs. Rationalizability

Every step of IBR reasoning adds another level of ToM reasoning and increments the depth of nested belief in rationality. When looking at things this way, IBR reasoning is strongly reminiscent of iterated dominance reasoning and rationalizability. Still, if we look more closely at these two solution con-

31. Jäger (2007) applies such a model to linguistic pragmatics.

$$
\begin{array}{lll}
\text{Base:} & \mathsf{S}_0 = \mathsf{S} & \\
& \mathsf{R}_0 = \mathsf{R} & \\
\text{Step:} & \mathsf{S}_{k+1} = \{s \in \mathsf{S}_n \quad | \quad \exists\, \rho \in \Pi_S \ : & \\
& \qquad \text{(RAT-S1)} \quad \rho \in \Delta(\mathsf{R}_k) & \\
& \qquad \text{(RAT-S2)} \quad s \in \mathrm{BR}(\rho)\ \} & \\
& \mathsf{R}_{k+1} = \{r \in \mathsf{R}_n \quad | \quad \exists\, \pi_R = \langle \Pr, \sigma, \mu \rangle \in \Pi_R \ : & \\
& \qquad \text{(RAT-R1)} \quad \sigma \in \Delta(\mathsf{S}_k) & \\
& \qquad \text{(RAT-R2)} \quad \pi_R \text{ is consistent} & \\
& \qquad \text{(RAT-R3)} \quad r \in \mathrm{BR}(\mu)\ \} & \\
\text{Limit:} & \mathrm{Rat} = \langle \bigcap_i \mathsf{S}_i, \bigcap_i \mathsf{R}_i \rangle &
\end{array}
$$

Figure 2.8: Standard (weak) rationalizability

cepts, we find interesting differences. I will suggest in this section that IBR should be regarded as a refinement of rationalizability, to which it adds focality of conventional meaning. More precisely even, the IBR model is a refinement of cautious rationalizability, which in turn is a refinement of strong rationalizability as defined by Battigalli (2006).

Weak Rationalizability

Recall from section 1.2.3 that a standard notion of rationalizability for signaling games is defined inductively as in figure 2.8. I will speak of WEAK RATIONALIZABILITY in the present context, because, firstly, we have already seen in section 1.2.3 that this solution concept is indeed very weak in the context of signaling games (cf. Zapater 1997; Battigalli 2006), and, secondly, we will look at two other notions subsequently, called strong and cautious rationalizability.

Let's first just superficially compare weak rationalizability to the basic IBR scaffolding in figure 2.6. There are three major differences to note. Firstly, whereas IBR allows for arbitrary restrictions of strategies in the base step, rationalizability begins with the full set of sender and receiver strategies. Secondly, rationalizability assumes a standard notion of best response in conditions (RAT-S1) and (RAT-R1), asking for some conjecture $\delta \in \Delta(X_n)$ about opponent strategies in X_n to rationalize behavior at level $(n+1)$. Here, my preferred version of IBR assumes unbiased beliefs $\delta = X_n$, while Jäger's IBR model would assume non-exclusive beliefs $\delta \in \Delta^+(X_n)$. Thirdly and lastly, the inductive step of rationalizability is purely eliminative, ruling out more and more strategies. In contrast to that, IBR rather defines a trajectory through the strategy space. This leads to a different assessment of the prediction of

the solution concepts in the limit. In finite games, rationalizability will always reach a fixed point, while IBR may loop, so that we need to collect all infinitely repeated strategies to define IBR's limit solution.

Despite these differences, it turns out that IBR is a rather conservative refinement of rationalizability in the sense that the limit prediction of the vanilla IBR model contains only rationalizable strategies.

Claim 2.4.1. Every solution of the vanilla IBR model is rationalizable.

This may as such not be too much of a surprise, given the weakness of standard rationalizability, but still it actually follows from an interesting further result, namely that the solution selected by the IBR model has the best response property.

BEST RESPONSE PROPERTY. Towards a proof of the above claim, let us define the notion of best response underlying weak rationalizability. Given a set of receiver strategies $\mathsf{R}' \subseteq \mathsf{R}$ define the set $\text{WBR}(\mathsf{R}') \subseteq \mathsf{S}$ as:

$$\text{WBR}(\mathsf{R}') = \left\{ s \in \mathsf{S} \mid \exists\, \rho \in \Delta(\mathsf{R}') \;:\; s \in \text{BR}(\rho) \right\}.$$

Similarly, for $\mathsf{S}' \subseteq \mathsf{S}$ let

$$\begin{aligned} \text{WBR}(\mathsf{S}') \;=\; \{ r \in \mathsf{R} \mid \exists\, \pi_R = \langle \Pr, \sigma, \mu \rangle \in \Pi_R \;:\; & \\ \sigma \in \Delta(\mathsf{S}') \;\wedge & \\ \pi_R \text{ is consistent} \;\wedge & \\ r \in \text{BR}(\mu) \}. & \end{aligned}$$

Following Pearce (1984), say that a pair of sets of sender and receiver strategies $\langle \mathsf{S}', \mathsf{R}' \rangle$ has the BEST RESPONSE PROPERTY (BR property for short) iff (i) $\mathsf{S}' \subseteq \text{WBR}(\mathsf{R}')$ and (ii) $\mathsf{R}' \subseteq \text{WBR}(\mathsf{S}')$. It is straightforward to show that:

Lemma 2.4.2. If $\langle \mathsf{S}', \mathsf{R}' \rangle$ has the BR property, then $\langle \mathsf{S}', \mathsf{R}' \rangle \subseteq$ Rat, i.e, all its strategies are rationalizable.

Proof. This is so, because all strategies in the sets S' and R' will be in the run for rationalizability at the outset of the iterated elimination procedure. Since, by the BR property, each set provides support for all the strategies in the other set, no strategy from either set will be eliminated during rationalizability. □

Lemma 2.4.3. The limit prediction IBR of the vanilla IBR model has the BR property.

	a_1	a_2	a_3	m_1	m_2	m_3
t_1	2,0	0,2	1,1	$\surd$	–	–
t_2	0,2	2,0	1,1	–	$\surd$	–
t_3	1,1	1,1	1,1	–	–	$\surd$

Figure 2.9: Game in which the IBR outcome is not CURB

Proof. Let IBR $= \langle \mathsf{S}^*, \mathsf{R}^* \rangle$. Then $s \in \mathsf{S}^*$ means that there is some subset $\mathsf{R}' \subseteq \mathsf{R}^*$ for which $s \in \mathrm{BR}(\mathsf{R}')$, i.e., for which s is a best response to an unbiased belief. But then $s \in \mathrm{WBR}(\mathsf{R}')$, and since $\mathrm{WBR}(\cdot)$ is actually monotonic, $\mathrm{WBR}(\mathsf{R}') \subseteq \mathrm{WBR}(\mathsf{R}^*)$, we find $s \in \mathrm{WBR}(\mathsf{R}^*)$, too. A similar argument applies for the receiver side. Taken together, then, IBR has the BR property and the above claim established. □

Claim 2.4.1 follows from these two lemmas.

DIGRESSION: CLOSURE UNDER BEST RESPONSE. In order to understand the solution of IBR and rationalizability even better, it may help to briefly digress and to define the converse notion to the BR property. So, let's say that a pair of strategies $\langle \mathsf{S}', \mathsf{R}' \rangle$ is CLOSED UNDER RATIONAL BEHAVIOR (or CURB for short) iff (i) $\mathrm{WBR}(\mathsf{R}') \subseteq \mathsf{S}'$ and (ii) $\mathrm{WBR}(\mathsf{S}') \subseteq \mathsf{R}'$ (see Basu and Weibull 1991, for more on CURB sets as solutions for games).

It transpires that IBR is not necessarily CURB. A game where this shows is given in figure 2.9. We could think of this game as a variation on the matching pennies game with cooperation option (figure 1.9). The IBR model predicts here that senders of type t_3 will announce the share option m_3 in the limit. Senders of type t_1 or t_2, on the other hand, will always have firm beliefs about their opponent's strategy and therefore always send a 'misleading message' m_1 or m_2. Similarly, the receiver's responses in the limit prediction always answer m_1 and m_2 with both a_1 and a_2:

$$\begin{aligned} \text{IBR} &= \langle \mathsf{S}^*, \mathsf{R}^* \rangle \\ &= \left\langle \left\{ \begin{array}{lcl} t_1 & \mapsto & m_1, m_2 \\ t_2 & \mapsto & m_1, m_2 \\ t_3 & \mapsto & m_3 \end{array} \right\}, \left\{ \begin{array}{lcl} m_1 & \mapsto & a_1, a_2 \\ m_2 & \mapsto & a_1, a_2 \\ m_3 & \mapsto & a_3 \end{array} \right\} \right\rangle \end{aligned}$$

However, for this game a sender strategy that sends m_3 in state t_1 or t_2 can also be a best response to R^*: IBR does not yield a CURB solution.[32]

32. This result is interesting in comparison to another recent proposal by Jäger (2008*b*):

Although being CURB is not necessarily always a desirable property for a selected solution, the example suggests that sometimes it might be appealing to try closing the IBR model's prediction under best responses. This could be a way of making sense of an agent's *Aha-Erlebnis*: after having reasoned herself through a non-trivial cycle, the agent realizes that any strategy occurring in IBR could be expected and plays a best response to some (biased or unbiased) belief in sets S^*, respectively R^*. This way we could think of extending the IBR sequence with a transfinite step to include level-ω players. I will leave this issue on this speculative note, because it is unclear to me what the proper logic of such reasoning should be, whether there is reasonable empirical evidence that could inform such transfinite modelling, and, last but not least, because this extension does not seem necessary for any linguistic application entertained in this thesis.

INTERMEDIATE SUMMARY. Since actually weak rationalizability is fairly weak, especially for cheap-talk signaling games, it is not a particularly striking result to find that IBR's limit prediction is rationalizable. It is, however, certainly non-trivial that IBR's limit solution has the BR property (lemma 2.4.3). This result tells us that IBR's solution is generally well-behaved, not only when proposition 2.2.1 applies to guarantee that IBR selects a PBE.

Still, the IBR model is more restricted than weak rationalizability. This is clear already by looking at, for instance, the some-all game, for which weak rationalizability does not discard any strategy profile, while IBR selects uniquely the intuitive scalar implicature play. This suggests that we could thus think of IBR as a refinement of rationalizability which integrates the impact of conventional meaning of cheap talk. This is indeed how I like to think of IBR. Yet, strictly speaking, the IBR model is more refined than weak rationalizability for *two* reasons, only one of which directly relates to semantic meaning.

Firstly, unlike rationalizability, IBR marks semantic meaning as a focal strategy at the outset of iteration of best responses. It is obvious that rationalizability cannot simply assume "semantic play" in its inductive base because its inductive step is purely eliminative: with honest sender S_0 and credulous receiver R_0 at the outset, rationalizability would never be able to account for

in this model, Jäger requires a solution to have both the BR property —actually, a *strong* BR property, based on strong rationalizability— and being a minimal CURB set that contains honest, truthful signaling of the sender. This conjunctive solution concept obviously does not guarantee existence of a solution in all games.

dishonest signaling and incredulous disbelief. In order to implement focal strategies, IBR has to be non-eliminative, with the consequence that the IBR sequence sometimes wanders infinitely through the strategy space. Still, at the end of the day, this seems just like the proper way of making rationalizability susceptible to conventional meaning as a focal strategy of unsophisticated agents.

Yet, secondly, the vanilla IBR model also places much stronger constraints on the belief formation of agents in each iteration step. Whereas weak rationalizability maintains all best responses to some belief $\delta \in \Delta(X_k)$ about opponent strategies from the set X_k, Jäger's model has non-exclusive beliefs $\delta \in \Delta^+(X_k)$ and the vanilla IBR model has unbiased beliefs $\delta = X_k$. These latter requirements of IBR are increasingly stronger and may lead to more strategies being discarded in the process of iteration. Additionally, this has noteworthy conceptual implications: IBR implicitly implements a further forward induction rationale, as its belief formation process is actually a strengthened version of that featured in *strong rationalizability*, to which we turn next.

Strong & Cautious Rationalizability

Strong rationalizability has been spelled out for arbitrary (but finite) dynamic games of incomplete information by Battigalli and Siniscalchi (2002) and worked out specifically for signaling games by Battigalli (2006). Different from the weak variety, strong rationalizability requires agents of level k to be firm in their belief that the opponent is of level $(k-1)$. The main idea is that even if an agent of level k holds an arbitrary conjecture about his opponent's play, she should try to rationalize all behavior, even surprise behavior, in a way compatible with her supposition that the opponent is of strategic level $(k-1)$. This effectively implements a forward induction rationale, as Battigalli and Siniscalchi argue. I will show in the following that IBR —my version and Jäger's— implicitly contains a similar, but strictly stronger concept of firm belief in opponent behavior and that thus both versions of IBR contain an additional element of forward induction reasoning (additional to a possible adoption of weak k-dominance or similar).

STRONG BELIEFS. To begin with, let us have a closer look at the strong version of rationalizability that Battigalli (2006) spells out for signaling games. The main conceptual difference between weak and strong rationalizability is that the latter requires in particular the receiver to rationalize any possible surprise

message in a way consistent with the assumed level of sophistication of the sender.[33] We will say that strong rationalizability requires the receiver at level k to have *strong belief* in sender behavior of level $(k-1)$. Basically the idea behind strong belief is this: take a set $\mathsf{S}' \subseteq \mathsf{S}$ of pure sender strategies; while in weak rationalizability, the receiver would then adopt some posterior belief consistent with a behavioral belief $\sigma \in \Delta(\mathsf{S}')$, strong belief in S' requires that if a message m gets used by *some* strategy $s \in \mathsf{S}'$, then the receiver should *not* adopt a posterior belief after m that considers it possible that m has been sent in a state different from the ones that would send m according to S'. Towards formalization, define

$$\mathsf{S}'(m) = \big\{t \in T \mid \exists s \in \mathsf{S}' \ : \ s(t) = m\big\}$$

as the set of states where message m could be sent in if the sender plays some arbitrary strategy in S'. With this say that a given receiver belief $\langle \Pr, \sigma, \mu \rangle$ satisfies STRONG BELIEF in S' iff

$$\mathsf{S}'(m) \neq \emptyset \Rightarrow \mu(\cdot|m) \in \Delta(\mathsf{S}'(m)).$$

Strong rationalizability is then defined just as weak rationalizability, only that condition (RAT-R1) in the inductive step for R_{k+1} is strengthened to (STR-RAT-R1) in order to require strong belief in S_k:[34]

$$\begin{aligned} \mathsf{R}_{k+1} = \{r \in \mathsf{R}_k \quad | \quad & \exists\ \pi_R = \langle \Pr, \sigma, \mu \rangle \in \Pi_R \ : \\ & \text{(STR-RAT-R1)} \quad \pi_R \text{ is a strong belief in } \mathsf{S}_k \\ & \text{(RAT-R3)} \qquad r \in \mathrm{BR}(\mu)\ \}. \end{aligned}$$

RUNNING EXAMPLE. Here is a simple example that shows what strong belief does, in comparison to the belief formation in weak rationalizability. (I will come back to this example repeatedly, because it also illustrates the differences between other relevant ways of forming beliefs.) Take a signaling game with just two states $T = \{t_1, t_2\}$ that are equally probable, and two messages $M = \{m_1, m_2\}$. Suppose that S_{run}, indexed for "running (example)," contains only

33. There are further differences between weak rationalizability and the solution presented by Battigalli, but none of these really matter for the present purposes. Most prominently, Battigalli also allows for diverging priors and arbitrary probabilistic beliefs about opponent behavior as fixed and unrevisable assumptions at the outset of rationalization.

34. Condition (RAT-R2) is superfluous given (STR-RAT-R1).

these two pure sender strategies:

$$s_1 = \left\{ \begin{array}{lcl} t_1 & \mapsto & m_1 \\ t_2 & \mapsto & m_1 \end{array} \right\}$$

$$s_2 = \left\{ \begin{array}{lcl} t_1 & \mapsto & m_1 \\ t_2 & \mapsto & m_2 \end{array} \right\}.$$

Under weak rationalizability's conditions on the receiver's belief formation, it is possible to obtain a posterior μ for which $\mu(t_1|m_2) > 0$. This is possible because under weak rationalizability the receiver may adopt the behavioral belief that the sender plays s_1 for sure. Under this belief, m_2 is a surprise message and beliefs with $\mu(t_1|m_2) > 0$ are consistent with this behavioral belief. This may be dubious because the only strategy that sends m_2 is s_2, and in this strategy m_2 is not sent in t_1. Accordingly, strong rationalizability restricts $\mu(\cdot|m_2)$ to the set $\{t_1\}$ of states where the message might plausibly get sent in S_{run}. This plausible restriction to strong beliefs protects the receiver, in a manner of speaking, from surprising himself, i.e., from adopting a too specific behavioral belief, so that surprise messages can be interpreted at random.

Still, the example also shows that strong belief is still rather lax in a different respect. For strong belief in the above set S_{run} does not rule out that the receiver forms a posterior μ for which $\mu(t_1|m_1) < \mu(t_2|m_1)$. Yet this is also rather peculiar, because there is no behavioral belief $\sigma \in \Delta(\mathsf{S}_{\text{run}})$ under which such posterior beliefs would be consistent. (Recall that we assumed a flat prior in this example.) This shows that strong belief is detached from any actual behavioral belief, except in restricting qualitatively that posteriors be formed in accordance with the sets $\mathsf{S}'(m)$.

This suggests that we should slightly strengthen strong belief somehow to respect reasonable behavioral beliefs. Happily, this is possible, but we need the concept of a *cautious belief*. This notion will be strictly stronger than strong belief. The concept of belief formation in IBR will turn out a special, strictly stronger case of cautious belief.

CAUTIOUS STRONG BELIEFS. Cautious belief is defined in terms of lexicographic beliefs. A LEXICOGRAPHIC BELIEF about events in set X is a (finite) vector $\langle \delta_1, \delta_2, \ldots, \delta_n \rangle$ of probability distributions $\delta_i \in \Delta(X)$.[35] The idea is that,

35. Cautious lexicographic beliefs have been studied in game theory, such as to give an epistemic characterization for iterated *weak* dominance (see Blume et al. 1991*a*,*b*). See Halpern (2009) and references therein for a recent overview on lexicographic beliefs and some comparison to other related belief representations.

roughly speaking, the probabilities assigned by δ_i are infinitely bigger than the probabilities assigned by δ_{i+1}. We could think of this as a hierarchy of beliefs.

The probability a lexicographic belief assigns to a single event $x \in X$ is then a vector $\langle \delta_i(x) \rangle_{i \leq n}$. When computing the agent's expected utilities of an action a given x based on these beliefs we similarly compute a vector with expected utilities $\langle \mathrm{EU}_i(a, x) \rangle_{i \leq n}$ for each δ_i. These expected utility vectors are then compared lexicographically:

$$\langle \mathrm{EU}_i(a, x) \rangle_{i \leq n} < \langle \mathrm{EU}_i(a', x) \rangle_{i \leq n} \quad \text{iff} \quad \mathrm{EU}_j(a, x) < \mathrm{EU}_j(a', x)$$

where $j = \min \{ i \leq n \mid \mathrm{EU}_i(a, x) \neq \mathrm{EU}_i(a', x) \}$. In a manner of speaking, lexicographic beliefs thus allow an agent's 'second guesses' to possibly influence her decisions, whenever her 'first guesses' leave her undecided.

In a sense, lexicographic beliefs could be taken to encode also counterfactual beliefs of agents. In particular, we can represent the situation where an agent assigns probability zero to an event $x \in X$, by setting $\delta_1(x) = 0$. But still, a lexicographic belief with this δ_1 may still give substantial information about what the agent would consider rational behavior if x actually occurred, because there could be some $j > 1$ for which $\delta_j(x) \neq 0$. In order to make sure that a lexicographic belief deals with *all* unexpected contingencies in this way we would like a belief to be *cautious*. Formally, say that a CAUTIOUS (LEXICOGRAPHIC) BELIEF $\langle \delta_1, \delta_2, \ldots, \delta_n \rangle$ in X has the property that for each $x \in X$ there is some j such that $\delta_j(x) \neq 0$.

How would we apply lexicographic beliefs to signaling games? Take a set $\mathsf{S}' \subseteq \mathsf{S}$ of sender strategies and let $\langle \sigma_i \rangle_{i \leq n}$ be a cautious lexicographic belief in S'. These *behavioral* lexicographic beliefs represent the receiver's conjectures —first, second, and so on— about sender behavior. For signaling games the interesting question becomes how this hierarchy of behavioral beliefs should be related to possible posterior beliefs of the receiver. For our modest purposes here, we should keep things simple and say that if the receiver's beliefs contain behavioral lexicographic beliefs, then $\langle \Pr, \langle \sigma_i \rangle_{i \leq n}, \mu \rangle$ is consistent iff

$$\mu(\cdot|m) = \frac{\Pr(t) \times \sigma_j(m|t)}{\sum_{t' \in T} \Pr(t') \times \sigma_j(m|t')}$$

where j is the smallest index for which m is not a surprise message under belief σ_j.[36]

36. If we wanted to be fussy here, we should actually define receiver beliefs with *partial*

If $\langle \sigma_i \rangle_{i \leq n}$ is a cautious lexicographic belief in S', then consistency defines the receiver's interpretation of all messages that he could expect to be sent by some type following a strategy in S'. Consistency, as defined here, furthermore confines the receiver's interpretation of all these messages as "consistently derived from" *some* conjecture σ_j about sender behavior in S', no matter how arbitrary and unrelated any of the conjectures in $\langle \sigma_i \rangle_{i \leq n}$ might be.

Cautious Rationalizability. These considerations already make it plausible that cautious beliefs help obtain a formulation of rationalizability that is slightly stronger than strong rationalizability as proposed by Battigalli (2006). Indeed, it is easy to see that consistent cautious belief entails strong belief:

Proposition 2.4.4. For every consistent and cautious belief $\langle \Pr, \langle \sigma_i \rangle_{i \leq n}, \mu \rangle$ in some set S', there is a strong belief $\langle \Pr, \sigma, \mu' \rangle$ in S' with $\mu = \mu'$.

Proof. To see this it suffices to note that if a message m is ever sent by some type under some strategy in S', then a consistent and cautious belief will have a smallest behavioral belief σ_j for which m is not a surprise message. Then, since $\mu(\cdot|m)$ is derived from this σ_j, it is guaranteed that $\mu(\cdot|m)$ has a support in the set of states that might send m according to σ_j and therefore $\mu(\cdot|m)$ has a support in $\mathsf{S}'(m)$. □

Our discussion of the above example already showed that the converse does not hold: not every strong belief has a corresponding consistent cautious belief. Cautious belief is a strictly stronger requirement on belief formation than strong belief.

We can then define cautious rationalizability by taking weak rationalizability with the exception that:

$$\begin{aligned} \mathsf{R}_{k+1} = \{ r \in \mathsf{R}_k \quad | \quad & \exists\, \pi_R = \langle \Pr, \langle \sigma_i \rangle_{i \leq n}, \mu \rangle \in \Pi_R \ : \\ & \text{(C-RAT-R1)} \quad \langle \sigma_i \rangle_{i \leq n} \text{ is a cautious belief in } \mathsf{S}_k \\ & \text{(RAT-R2)} \quad \pi_R \text{ is consistent} \\ & \text{(RAT-R3)} \quad r \in \mathrm{BR}(\mu) \ \}. \end{aligned}$$

lexicographic posteriors $\langle \mu_i \rangle_{i \leq n}$ such that $\mu_j(\cdot|m)$ is consistent with Pr and σ_j whenever m is not a surprise message under belief σ_j, and undefined otherwise. Expected utility would then be defined as usual under lexicographic beliefs, so that whenever a message m surprises the receiver under belief σ_1, we would keep looking for some integer j where m is not a surprise message for expected utility comparison.

Clearly, cautious rationalizability is a further strengthening of the forward induction rationale in strong rationalizability. Just as the latter, cautious rationalizability requires the beliefs of the receiver at level k that the sender is of level $(k-1)$ to be firm. Still, cautious rationalizability further strengthens this idea by requiring that firm beliefs that the sender is of level $(k-1)$ are to be derived, by consistency, from some behavioral belief in level $(k-1)$ compatible play.

IBR Strengthens Cautious Belief. The upshot of this discussion is that the belief formation of IBR is to be regarded as a further refinement of cautious belief. If $\langle \Pr, \sigma, \mu \rangle$ is a non-exclusive receiver belief in set S', then $\langle \Pr, \langle \sigma \rangle, \mu \rangle$ is a consistent cautious belief in S'. The converse, however, does not hold. Consistent cautious beliefs in some set can contain posteriors that cannot be reached by a consistent non-exclusive belief. Again, our running example from above illustrates this: under the assumption that priors are flat, it is not possible to have a consistent non-biased belief in the set S_{run} for which $\mu(t_1|m_1) = \mu(t_2|m_1)$, but it is possible to have this posterior under a consistent cautious belief.

Since moreover unbiased beliefs are clearly a special case of non-exclusive beliefs, we find the following strict hierarchy of belief formation requirements:

$$\text{strong} \supset \text{cautious} \supset \text{non-exclusive} \supset \text{unbiased}$$

The IBR models in terms of non-exclusive or unbiased beliefs come out as further conservative refinements of cautious rationalizability where *both* sender *and* receiver of level k obey the forward induction rationale of a firm belief in level $(k-1)$ behavior.

We could then think of non-exclusive and unbiased beliefs as a forward induction strategy in the sense that these requirements *prevent* counterfactual beliefs wherever possible. From this point of view, it is also clear why this is a forward induction rationale different from that of weak k-dominance. Firm belief requires the receiver of level k to rationalize each message as if sent by a level-$(k-1)$ sender as much as possible. If that is really not possible because no level-$(k-1)$ sender would ever send a given message m, then and only then could weak k-dominance kick in and require that m be rationalized as sent by a level-$(k+1)$ sender. Taken together these two forward induction procedures imply that the receiver has a firm belief that the sender is *at least* of level $(k-1)$.

SUMMARY. To wrap up, the close comparison with rationalizability suggests that we should look at IBR as a solution concept that selects rational behavior consistent with common belief in (i) rationality, (ii) focal conventional meaning, (iii) non-exclusive, or even unbiased beliefs, and optionally also (iv) truth ceteris paribus, and (v) weak k-dominance. Whether this conjecture holds to scrutiny under a formally spelled out epistemic characterization result is an interesting open question for further research.

2.5 Semantic Meaning and Credibility

In order to round off the exposition of the IBR model, we should check whether the model actually fulfills its intended purpose. The IBR model is intended as an alternative solution concept for use in GTP that suitably integrates conventional meaning into cheap talk games. The IBR model implements conventional meaning as a focality restriction on reasoners' belief formation. In order to check whether this is an appropriate implementation of semantic information, we should formulate a notion of *message credibility* that tells us under which strategic circumstances —think: payoff constellations, available messages and their semantic meaning— trust in conventional meaning is warranted. Matching this against intuition when to trust the semantic meaning of a message, we may thus test the proposed solution concept. This is what the following section 2.5.1 does.

Subsequently, section 2.5.2 compares the present approach to the two most influential approaches to refining game theoretic solutions by conventional meaning, namely Joseph Farrell's *neologism-proofness* as a refinement of equilibrium (Farrell 1993), and Rabin's *credible message rationalizability* as a refinement of rationalizability (Rabin 1990). There is a major difference between these latter approaches and the IBR model. While the former have tried to define credibility, as it were, from the outside in order to feed an abstract notion of credibility into an existing solution concept as a refinement, the IBR model feeds semantic information directly into a novel solution concept and has a credibility notion fall out. I argue in this section that the game theorist's approach is too inflexible, because it does not apply to especially those games that we are interested in for pragmatic applications.[37]

37. Although this section will focus on the work of Farrell and Rabin, similar criticism applies to improvements of Farrell's approach (e.g. Myerson 1989; Matthews et al. 1991) as well as Rabin's (e.g. Zapater 1997).

2.5.1 Message Credibility

Section 1.2.4 has introduced the intuitive notion of message credibility in a cheap talk signaling game. Clearly, in a game of pure coordination, such as the wine-choice scenario from section 1.2.2, there is no reason at all to suspect that messages could be used untruthfully. But if we look at the matching pennies game with cooperation option given in figure 1.9 on page 39, we feel that only message $m_{\texttt{coop}}$ is credible while $m_{\texttt{heads}}$ and $m_{\texttt{tails}}$ are not. Intuitively, we expect something like the play given in figure 1.10a on page 40 (and not, for instance, the play in figure 1.10b).

The IBR model solves this case straightforwardly and allows a very accessible definition of message credibility on top of it. To see this, let's briefly just calculate the model's predictions for the matching pennies game in figure 1.9. As we have $R_0 = R_1$, it suffices to look at the R_0-sequence. The naïve receiver R_0 is just credulous and takes all messages to be literally true:

$$R_0 = \left\{ \begin{array}{lcl} m_{\texttt{heads}} & \mapsto & t_{\texttt{heads}} \\ m_{\texttt{tails}} & \mapsto & t_{\texttt{tails}} \\ m_{\texttt{coop}} & \mapsto & t_{\texttt{coop}} \end{array} \right\}.$$

Based on such a credulous receiver strategy, S_1 will send $m_{\texttt{coop}}$ in $t_{\texttt{coop}}$, which is true in this state. But in states $t_{\texttt{heads}}$ and $t_{\texttt{tails}}$ the sender has an incentive to send false messages. Since S_1 believes that the receiver plays according to R_0, in order to maximize her expected utility in state $t_{\texttt{heads}}$, for instance, she will send $m_{\texttt{tails}}$, although this message is false in this state:

$$S_1 = \left\{ \begin{array}{lcl} t_{\texttt{heads}} & \mapsto & m_{\texttt{tails}} \\ t_{\texttt{tails}} & \mapsto & m_{\texttt{heads}} \\ t_{\texttt{coop}} & \mapsto & m_{\texttt{coop}} \end{array} \right\}.$$

If the receiver believes that the sender plays S_1, he will, in a manner of speaking, undo the reversal of messages and best respond with:

$$R_2 = \left\{ \begin{array}{lcl} m_{\texttt{heads}} & \mapsto & t_{\texttt{tails}} \\ m_{\texttt{tails}} & \mapsto & t_{\texttt{heads}} \\ m_{\texttt{coop}} & \mapsto & t_{\texttt{coop}} \end{array} \right\}.$$

The sender, in turn, will then best respond with an entirely truthful sending strategy:

$$S_3 = \left\{ \begin{array}{lcl} t_{\texttt{heads}} & \mapsto & m_{\texttt{heads}} \\ t_{\texttt{tails}} & \mapsto & m_{\texttt{tails}} \\ t_{\texttt{coop}} & \mapsto & m_{\texttt{coop}} \end{array} \right\}.$$

The receiver's best response to this is again the strategy R_0 and so the sequence starts to loop. The IBR model then predicts that any of the above strategies may be played under common belief in rationality. Crucially, this means that, as we would intuitively expect, the message $m_{\texttt{coop}}$ faithfully signals the unique state where it is true, but the messages $m_{\texttt{heads}}$ and $m_{\texttt{tails}}$ do not. In accordance with intuition, we could say that message $m_{\texttt{coop}}$ is credible, because it is used truthfully by every sender type in the whole IBR sequence. Messages $m_{\texttt{heads}}$ and $m_{\texttt{tails}}$ are not, because there are sender types, such as S_1, who use these messages untruthfully.

Defining Credibility. Extrapolating from this example, let us define that a message m is CREDIBLE iff there is no sender type in the whole IBR sequence that uses m in a state where it is not true. Formally, m is credible iff there are no k and t such that $m \in S_k(t)$ and $t \notin [\![m]\!]$. This notion then captures whether there is ever any positive incentive for the sender to send a false message.

In a manner of speaking, message credibility tests whether the boundary of conventional meaning might ever be crossed in a particular situation due to the speaker's interests. Yet, it needs to be pointed out for clarity that message credibility does *not* capture whether the sender tries to, in intuitive terms, mislead, trick or deceive the receiver. Take, for instance, the sender type S_1 in the above matching pennies example. Sender S_1 believes that the receiver plays according to R_0. Given the sender's preferences and these beliefs, it is feasible to say that S_1 expects to be able to *mislead* the receiver:[38] she sends $m_{\texttt{heads}}$ in $t_{\texttt{tails}}$, for instance, expecting to induce a false belief with a false message. However, this kind of misleading behavior is *not* what is crucial for our definition of message credibility. To see this we simply need to look at sender type S_3 in the same example who could also be said to try to mislead R_2. In the latter case, S_3 believes that the receiver plays according to R_2 and so sends $m_{\texttt{heads}}$, for instance, in $t_{\texttt{heads}}$ expecting to induce a false belief with a true message. For our definition of message credibility only the former case counts; attempts to 'mislead with the truth,' so to speak, do not count for message credibility which is solely concerned with assessing violations of conventional meaning, but not with violations of receiver expectations.

Let me briefly enlarge on this for clarity by giving two further illustrating examples — one in which a false and formally incredible message is intu-

38. Since S_1 has these beliefs and preferences, we can maybe even say that she *wants* or even *intends* to mislead the receiver in this case. However, strictly speaking, intentions are not formally represented in the game, only preferences and beliefs are.

	a_1	a_{23}	m_{12}	m_3
t_1	1,1	0,0	$\surd$	–
t_2	0,0	1,1	$\surd$	–
t_3	0,0	1,1	-	$\surd$

Figure 2.10: White lie

	$a_{\exists\neg\forall}$	$a_{\forall}$	$m_{\texttt{some}}$	$m_{\texttt{all}}$
$t_{\exists\neg\forall}$	1, 1	0, 0	$\surd$	–
$t_{\forall}$	1, 0	0, 1	$\surd$	$\surd$

Figure 2.11: Misleading implicature

itively not misleading; and one in which a true and formally credible message is intuitively misleading.

White Lies. Firstly, consider the game in figure 2.10. Here, the best response of the sender S_1 to a credulous receiver R_0, who interprets messages literally, is to send m_3 in state t_2 in order to induce action a_{23}. Although this message is false in this state, we may feel that this is not a genuine case of malicious misleading, as it is for the receiver's benefit, too.[39] The message m_3 is thus incredible, because it is used untruthfully along the IBR sequence, but it is not (maliciously) misleading to the receiver's detriment. Speaking the (only available) truth in t_2 could simply induce the wrong action for this state.

Misleading Implicatures. Secondly, consider the game in figure 2.11 which is very much like a scalar implicature case, but where preferences are no longer perfectly aligned. This game models a dialogue situation like (19) in which (it is common knowledge that) Riko would not want to concede whether all of her friends are Buddhists: whether some or all of her friends are Buddhists, Riko wants Saki to believe that only some of her friends are; Saki, on the other hand, wants to know the truth.

(19) a. Saki: How many of your friends are Buddhists?
b. Riko: Some of them are.

In the IBR model we find a sender type, namely S_2 who intuitively misleads the receiver R_1 by, what we could call, abusing the scalar implicature inference. To see this, notice that, despite the non-aligned preferences, R_1 does not

39. We could think of this case —with some due abstraction— as a 'white lie' in order to save face: a wife asks "Do these shoes make my … look big?" and a kind husband, knowing that the question is really about whether the shoes are affordable (a_1) or not (a_{23}), may answer "No, of course not." To make sense of the abstract example along these lines we need to imagine that the husband cannot say anything else than "yes" and "no", and that the wife does not really want to know the true and honest answer to her actual question.

	$\Pr(t)$	a_1	a_2	a_3	m_{12}	m_{23}
t_1	$1/8$	1,1	0,0	0,0	$\surd$	–
t_2	$3/4$	0,0	1,1	0,0	$\surd$	$\surd$
t_2	$1/8$	0,0	0,0	1,1	–	$\surd$

Figure 2.12: Some game without a name

consider the sender strategically sophisticated and thus plays:

$$R_1 = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_{\exists\neg\forall} \\ m_{\texttt{all}} & \mapsto & t_{\forall} \end{array} \right\}.$$

Based on this receiver strategy, S_2 will send message $m_{\texttt{some}}$ in state $t_\forall$ in order to induce a false belief in state $t_{\exists\neg\forall}$. This is to the receiver's detriment, and thus a case of intuitive deception, although semantically the message is true and overall credible.[40]

The Role of TCP in Credibility. The last two examples made clear that the above notion of message credibility is not about whether the sender misleads or morally wrongs the receiver, but only about whether the sender ever has an incentive to deviate from semantic meaning. If all situated reasoning takes place within the conventional meaning of a message, the message is credible in this formal sense; otherwise it is incredible. Whether the sender deviates from semantic meaning or not in a situation where she is otherwise indifferent, i.e., whether the TCP assumption is in place or not, will thus also clearly affect whether a message is deemed credible or not. Here is a game that shows this.

Take the example in figure 2.12. The non-flat priors in this game have the receiver R_0 respond to both available messages in the same way:

$$R_0 = \left\{ \begin{array}{lcl} m_{12} & \mapsto & a_2 \\ m_{23} & \mapsto & a_2 \end{array} \right\}.$$

But that means that S_1 is actually indifferent between sending a true and an untrue message in states t_1 and t_3, because she expects a utility of zero from

40. The example also shows how the model predicts scalar implicatures to vary with the degree of preference alignment. It is well known that implicatures are upper-bounded, so to speak, by the extent to which the speaker is cooperatively adopting the hearer's interests (cf. O'Hair 1969; Green 1995; Sperber and Wilson 1995; Carston 1998). The benefit of GTP is that it can make various degrees of conflict explicit (see Franke et al. to appear).

sending either message m_{12} or m_{23} in either state. Without TCP assumption, we would therefore have to say that all messages are incredible in this scenario, because there are sender types who would use false messages simply because they are indifferent between truth and falsity as far as expected utilities are concerned. This predicament seems a bit harsh. But under the TCP restriction, the same definition of message credibility yields that all messages in the game in figure 2.12 are credible, which may be a more intuitive verdict.

Truth of Surprise Messages. It is worthwhile mentioning that the TCP assumption as such does not imply that the receiver will interpret surprise messages as true (Jäger 2008*c*, makes this assumption). With or without the forward induction principle of weak k-dominance, the receiver can adhere to the TCP expectation, and still believe that a surprise message was possibly sent in a state where it was not true. In fact, no version of the IBR model, with or without either TCP and forward induction, constrains the receiver's counterfactual beliefs in any way to acknowledge conventional meaning.

This may be surprising, because we may assume that the receiver's counterfactual beliefs should also at least be informed by conventional meaning. Maybe we would even be inclined to say that weak k-dominance applies only to the set of states where a given surprise message was true to begin with, and that only when no state satisfies both the presumption of truth and the requirements of forward induction would the receiver drop the presumption of truth. This would correspond to general principles of belief revision in the sense that the receiver holds on to the assumption that a message is true and rationally used, for as long as this is possible, and would drop the truth presumption only where this would contradict rationality.

These are all fair considerations in theory, but interestingly it turns out that none of this is necessary to make the IBR model do its work, at least not for any of the examples and problems considered in this thesis. This does not mean that there may not be other cases for which the receiver's counterfactual beliefs should be constrained to respect conventional meaning. It simply means that we may leave this question as an open issue for further research *if and where* needed for reasonable applications.

Credible Surprises. Since message credibility is defined here in terms of sender behavior this also means that surprise messages are not necessarily deemed incredible, even though the receiver's counterfactual beliefs are not restricted by semantic meaning. In other words, a message that surprises

some receiver type can come out credible or incredible, depending on whether the sender ever actually sends that message in a state where it is not true. This is as it should be, except perhaps for the fact that the current notion deems all messages that are *never* sent along the IBR sequences credible. To this it may be objected that my favored notion does not dissect far enough into any possible and rational 'counterfactual sending behavior': it does not address whether the sender *would* ever use an unused message untruthfully within the limits of rationality. I am not aware of any conceptual problem or example that would require such an extension, and leave this issue for further research to those who find it problematic.

2.5.2 Credibility-Based Refinements

Let us finally compare IBR's notion of credibility with the seminal game theoretic work on this issue, namely Farrell's (1993) *neologism-proofness* and Rabin's (1990) *credible message rationalizability*.

Neologism-Proofness

The classic idea on message credibility is due to Farrell (1993).[41] According to Farrell, a message m is credible if, roughly, all the sender types in $[\![m]\!]$ want the receiver to believe that m is true while no type outside of $[\![m]\!]$ wants it believed (cf. Farrell and Rabin 1996, p. 106). Based on this notion of credibility, Farrell (1993) defines a refinement of equilibrium for signaling games, called *neologism-proofness*, that aims to rule out unintuitive equilibria whenever these conflict with the conventional meaning of a shared and highly expressive language. Since Farrell's notion of credibility is used to rule out unintuitive equilibria, credibility is actually defined with respect to a given reference equilibrium. Here is the idea in formal terms.

Take a cheap talk signaling game and let $\langle \sigma, \rho, \mu \rangle$ be a perfect Bayesian equilibrium of this game. Let us write $U_S^*(t)$ as the expected utility outcome of the sender in state t under the equilibrium play. A message m is a CREDIBLE NEOLOGISM with respect to the PBE $\langle \sigma, \rho, \mu \rangle$ iff (i) m is a surprise message given σ, (ii) all and only the types $t \in [\![m]\!]$ benefit from a literal interpreta-

41. Small history note: Farrell's paper had been in circulation for years before it got published. Whence that, for instance, already Myerson (1989) and Rabin (1990) are fully spelled out and published reactions to Farrell's original ideas.

tion of m, i.e., $\{t \in T \mid \mathrm{EU}_S(m, t, R_0) > \mathrm{U}^*_S(t)\} = [\![m]\!]$.[42] A NEOLOGISM-PROOF equilibrium is a PBE for which there are no credible neologisms.

CRITIQUE OF NEOLOGISM-PROOFNESS. Several things are noteworthy about this refinement notion. First of all, in order to make neologism-proofness bite, we have to buy into a RICH LANGUAGE ASSUMPTION: Farrell assumes that for every PBE of a given signaling game and for every subset $X \subseteq T$ of states there is a message m in the game with $[\![m]\!] = X$ such that m is a neologism, i.e., not used in the PBE. This strong requirement on expressibility in the set of messages of a signaling game renders neologism-proofness *as it stands* inapplicable to cases of interest for linguistic pragmatics where strengthening of the semantic meaning of a message arises from (partial) inexpressibility (at equal cost). If we want to apply neologism-proofness to, for instance, the some-all game so as to select the intuitive scalar implicature play, we need to give up the rich language assumption and amend neologism-proofness accordingly (see below).

Secondly, whether or not sufficient messages are available, neologism-proofness will always only help rule out pooling equilibria, but not as such rule out fully revealing communication with (partially) untrue but cheap signalling. The strategy profile number 16 in figure 1.5 section 1.2.2, for instance, cannot be ruled out by neologism-proofness alone. This is so because basically all sender types can already induce their most preferred receiver response. It then does not really matter that they (partly) do so with a false message. Thus conceived, neologism-proofness falls short of implementing semantic meaning in the solution concept *tout court*.

Lastly, what is worse, not even necessarily all unintuitive pooling equilibria are ruled out by neologism-proofness in the most satisfactory way (contra van Rooij 2008). Take the persistent pooling PBE number 10 in figure 1.5. If the set of messages in the signaling game is restricted to $m_{\texttt{some}}$ and $m_{\texttt{all}}$, then this PBE is neologism-proof because the only neologism $m_{\texttt{all}}$ is not credible since $t_{\texttt{all}}$ does not *strictly* benefit from sending it and having it believed. We would need a different, equally costly neologism $m_{\texttt{sbna}}$ in order to rule out this pooling equilibrium. But that actually means that we are playing a different game to begin with and also that we get other PBEs, such as the one using $m_{\texttt{sbna}}$ to express $t_{\exists\neg\forall}$ and $m_{\texttt{some}}$ to express $t_{\forall}$, that need to explained away.

42. Notice that I am (ab)using notation of the IBR model here: to represent the sender's expectation under a literal interpretation, I use R_0 as a sender belief, although strictly speaking, of course, Farrell's approach does not as such involve IBR-style reasoning or IBR types.

	a_1	a_2	a_3	m_1	m_2
t_1	4,3	3,0	1,2	$\surd$	$-$
t_2	3,0	4,3	1,2	$-$	$\surd$

Figure 2.13: "Best message counts"

Farrell Credibility. But even though neologism-proofness does not suit our needs as it stands, it is nonetheless worthwhile to excavate the underlying idea of credibility for further comparison to the IBR model. The first thing that strikes us, of course, is that neologism-proofness is tied to neologisms. The main idea of *Farrell-credibility* is rather this:

(20) Given a PBE as reference, a message m is Farrell-credible iff all and only types in $[\![m]\!]$ benefit from a literal interpretation of m, i.e., $\emptyset \neq \{t \in T \mid \mathrm{EU}_S(m, t, R_0) > \mathrm{U}^*_S(t)\} = [\![m]\!]$.

As we have seen above already, this notion heavily relies on the rich language assumption that every possible subset of states is expressible with a neologism. Yet, once we allow for limited expressiveness and pragmatic strengthening, it should also be possible for a message to be pragmatically enriched and still be credible. Take, for instance, the scalar inference from $m_{\texttt{some}}$ to interpretation $t_{\exists\neg\forall}$. Does ruling out $t_\forall$ from the interpretation of $m_{\texttt{some}}$ mean that this message is not credible? I would not think so. Hence, if we give up the rich language assumption, thus allowing for pragmatic enrichment to be consistent with credibility, we should maybe adapt Farrell's notion slightly:

(21) Given a PBE as reference, a message m is Farrell-credible iff some and only types in $[\![m]\!]$ benefit from a literal interpretation of m, i.e., $\emptyset \neq \{t \in T \mid \mathrm{EU}_S(m, t, R_0) > \mathrm{U}^*_S(t)\} \subseteq [\![m]\!]$.

This permits, e.g., $m_{\texttt{some}}$ to be credible under scalar enrichment.

Still, there are other problems with Farrell's notion of credibility and, by extension, the equilibrium refinement that it induces. Matthews et al. (1991) criticize Farrell-credibility as being *too strong*. Their argument centrally builds on the example in figure 2.13. Compared to the babbling equilibrium, in which R performs a_3, messages m_1 and m_2 are intuitively credible: both S^{t_1}, as well as S^{t_2} have good reasons to send m_1 and m_2 respectively. Communication seems possible and utterly plausible. However, neither message is Farrell-credible, because for $i, j \in \{1, 2\}$ and $i \neq j$ not only senders of type t_j, but also of type t_i prefer the receiver to play a best response to a literal interpretation

of m_j, which would trigger action a_j, over the no-communication outcome a_3.

The problem with Farrell's notion is obviously that just doing better than equilibrium is not enough reason to send a message, when sending another message is *even better* for the sender: when evaluating the credibility of a message m, we have to take into account *alternative forms* that $t \notin [\![m]\!]$ might want to send (under the assumption that all of the forms not occurring in equilibrium are interpreted *literally*). This is also the line that Matthews et al. (1991) take when they seek to define which sets of signaling strategies as a whole can be deemed credible, by maintaining the gist of Farrell-credibility and lifting equilibrium to sets of strategy profiles. However, a different way of solving the problem with Farrell's notion suggests itself from an IBR perspective. What seems to be at stake is that the receiver wonders: "which types of senders would send this message given that I believe it *literally*?" This suggests that, at least at this stage of sophistication, whether a message is credible or not depends on whether it is rational for the sender of type S_1 —who, after all, believes in literal uptake of messages— to send messages untruthfully.

STALNAKER-CREDIBILITY. Indeed, such a version of credibility could be read off the proposal of Stalnaker (2006). Stalnaker gives a very general notion of *prima facie rationality* which he informally states as follows:

> "A message m for S of type t is *prima facie rational* if and only if the expected value, for S, of sending m, and having it believed, is at least as great as the expected value of sending any alternative message."
>
> (Stalnaker 2006, p. 97)

Based on this notion of prima facie rationality, Stalnaker offers a straightforward definition of credibility, namely:

> "A message is credible iff it is [prima facie] rational for some types, and only for types for which it is true." (Stalnaker 2006, p. 93)

Stalnaker then suggest, still rather informally, that (something like) this notion of credibility should be integrated as a constraint on receiver beliefs —believe a message iff it is credible— into an epistemic model of the game together with some appropriate assumption of (common) belief in rationality. The class of game models that satisfies rationality and credibility constraints would then ultimately define how signals are used and interpreted.

Unfortunately, due to the preliminary if not speculative nature of Stalnaker's proposal, it is not entirely clear what the phrase "having it believed"

	a_1	a_2	a_3	a_4	m_{12}	m_{23}	m_{13}
t_1	4,5	5,4	0,0	1,4	$\surd$	–	$\surd$
t_2	0,0	4,5	5,4	1,4	$\surd$	$\surd$	–
t_3	5,4	0,0	4,5	1,4	–	$\surd$	$\surd$

Figure 2.14: "Further iteration"

in the definition of prima facie rationality means. It is plain to see, however, that if it is to mean "having it believed *literally*", then translating the notion into the present IBR terminology, prima facie rationality comes down to simple Bayesian rationality for the sender given a belief in receiver behavior R_0. Under this reading, a notion of *Stalnaker-credibility* is the following:[43]

(22) A message m is STALNAKER-CREDIBLE if and only if

$$\emptyset \subset \left\{ t \in T \mid \forall m' \in M \ : \ \mathrm{EU}_S(m, t, R_0) \geq \mathrm{EU}_S(m', t, R_0) \right\} \subseteq [\![m]\!] \,.$$

Indeed, Stalnaker credibility seems fairly intuitive and matches our intuitions in many cases. It is plain to see also that it is, loosely speaking, a part of my preferred notion of credibility. Still, Stalnaker credibility, as spelled out here, is in a sense self-refuting, as the following example in figure 2.14 from Myerson (1989) and Matthews et al. (1991) shows. In this game, all the available messages m_{12}, m_{23} and m_{13} are Stalnaker-credible, because if R interprets literally S will use message m_{12} in state t_1, message m_{23} in state t_2, and m_{13} in state t_3. No message is used untruthfully by any type. However, if R realizes that exactly S^{t_1} uses message m_{12}, he would rather not play a_2, but a_1. But if the sender realizes that message m_{12} triggers the receiver to play a_1, suddenly sender type t_3 wants to send m_{12} *untruthfully*. This example shows that Stalnaker-credibility is a reliable start, but stops too short. If messages are deemed credible and therefore believed, this may create an incentive to mislead. This is why the notion of credibility that I gave is based on the full IBR sequence, not just at the starting junction between R_0 and S_1.

Credible Message Rationalizability

Rabin (1990) responds to Farrell's credibility-based equilibrium refinement, noting critically among other things that neologism-proofness does not guar-

43. I should remark that I am not aiming for a precise Stalnaker exegesis here and that the label "Stalnaker-credibility" is not meant to discredit Stalnaker's work, but rather to credit it, although I argue that the notion bearing this name has shortcomings.

antee existence of a solution. Instead, Rabin therefore offers a non-equilibirum, rationalizability-based account of credible cheap talk that does guarantee existence.[44] Rabin's account, call it *credible message rationalizability* or CMR for short, is superficially very similar to the IBR model and it therefore pays to scrutinize Rabin's proposal and to compare CMR to IBR.

At the heart of CMR is a notion of message credibility as a self-signaling message that will be believed and exclusively used by all types for which it is true. In vague terms, we could say that 'play maximally consistent with credibility' is then fed into standard rationalizability as an initial restriction of the strategy space. The outcome of iterated strict dominance, thus restricted, is then the prediction of CMR. (Notice the close resemblance to Stalnaker (2006).) To understand Rabin's proposal, let us first look at CMR's notion of credibility and then at how it is used to restrict rationalizability.[45]

RABIN-CREDIBILITY. The definition of credibility that underlies CMR looks suspiciously like an amended version of the R_0-S_1-R_2 part of the IBR model. In first approximation, we may say that Rabin takes a message m to be credible if (i) all types in $[\![m]\!]$ obtain their best payoff when R interprets the statement literally, and (ii) R's optimal response to m does not change when he takes into account that certain types outside of $[\![m]\!]$ might also make the statement if literally interpreted.

Working towards a precise, formal definition of Rabin-credibility, let A^* be the set of all *zero-order rationalizable actions*, i.e. actions that are optimal for some given belief of the receiver:[46]

$$A^* = \left\{ a^* \in A \mid \exists \delta \in \Delta(T) \ : \ a^* \in \arg\max_{a \in A} \sum_{t \in T} \delta(t) \times \mathrm{U}_R(t, a) \right\}.$$

The actions in A^* provide an upper bound on what the sender can hope to be able to induce as a response in a rational receiver. We could thus speak of *inducible actions*.

44. Myerson (1989), for instance, takes issue with the fact that neologism-proofness may rule out all equilibria in some games, and consequently shows how to overcome this non-existence problem using Aumann's (1974) notion of *correlated equilibrium*.

45. The exposition here is a slightly simplified distillation of Rabin's account and takes into account the corrected version of CMR given by Rabin (1992).

46. For cheap talk games the zero-order rationalizable actions are the same for each message, obviously. We can also leave out specification of the message in the utilities of sender and receiver.

According to Rabin, a type $t \notin [\![m]\!]$ might want to mislead with message m unless (i) literal uptake would always trigger exactly her worst inducible outcomes, or (ii) there is a message which is better than m under a literal interpretation. Formally, Rabin defines $Y^*(m)$ as the set of all types that may want to mislead with message m as:[47]

$$\begin{aligned} Y^*(m) \;=\; \{t \notin [\![m]\!] \mid R_0(m) \neq \arg\min_{a \in A^*} U_S(a,t) \quad \text{or} \\ \neg\exists m' \; \forall a \in R_0(m) \; \forall a' \in R_0(m') \; : \; U_S(t,a) < U_S(t,a')\}. \end{aligned}$$

When interpreting a message m, the receiver may realize that $Y^*(m)$ is not empty, i.e., that some types might want to mislead with message m. He should then perhaps not take message m literally and make room for the possibility that types in $Y^*(m)$ have sent this message. The set of posterior beliefs that R might want to adopt if he takes possible misleadingness of m (as defined by $Y^*(m)$) into account is the set $\Pi(m)$ of probability distributions (not scaled to 1 for convenience) such that $\pi \in \Pi(m)$ whenever:

$$\pi(t) = \begin{cases} \Pr(t) & \text{if } \; t \in [\![m]\!] \\ p \in [0; \Pr(t)] & \text{if } \; t \in Y^*(m) \\ 0 & \text{otherwise.} \end{cases}$$

With all this we can define *Rabin-credibility* as follows: a message m is Rabin-credible iff

(i) for all $t \in [\![m]\!]$, $R_0(m) = \arg\max_{a \in A^*} \mathrm{U}_S(t,a)$ and

(ii) for all $\pi \in \Pi(m)$, $R_0(m) = \arg\max_{a \in A^*} \sum_{t \in T} \pi(t) \times \mathrm{U}_R(t,a)$.

In words, a message m is Rabin-credible if (i) m induces, when taken literally, exactly the set of all sender-best actions (from the set of inducible actions) of all types in $t \in [\![m]\!]$ and (ii) the set of best actions, according to a literal interpretation of m, is exactly the same as the set of best actions under any belief that places some positive probability also on those types that are not in $[\![m]\!]$, but might conceivably want to mislead with this message.

Credible Message Rationalizability. Based on this notion of credibility we would like to define sender and receiver play that is, in a manner of speaking, maximally consistent with credibility. Towards this end, let us suppose

47. I am again using R_0 as notation for a rational response to a literal interpretation.

for simplicity that there is exactly one message m for every non-empty subset of states $T' \subseteq T$ such that $[\![m]\!] = T'$.[48] We would then like to look at a set M' of credible messages whose semantic meaning yields a partition of a subset $T' \subseteq T$ of sender types, i.e., we look at sets M' of credible messages for which the set

$$\{X \subseteq T \mid \exists m \in M' \ : \ X = [\![m]\!]\}$$

is a partition.[49] If M' is such a set we can define the initial restriction of rationalizability as follows:

$$\begin{aligned} \mathsf{S}_0 &= \{s \in \mathsf{S} \mid \forall t \in T \, \forall m \in M' \ : \ (t \in [\![m]\!] \rightarrow s(t) = m \wedge \\ &\qquad t \notin [\![m]\!] \cup Y^*(m) \rightarrow s(t) \neq m)\} \\ \mathsf{R}_0 &= \{r \in \mathsf{R} \mid \forall m \in M' \ : \ r(m) \in R_0(m) \text{ and} \\ &\qquad \forall m \in M \ : \ r(m) \in A^*\} \end{aligned}$$

The further inductive steps follow standard rationalizability.

CMR vs. IBR. The first thing that might strike us about CMR is that it is fairly complicated. This speaks for IBR: if predictions were otherwise the same, we should prefer the perspicuity of IBR over CMR. Yet, predictions are not the same; in fact, CMR is fairly weak, so much so that it is inapplicable to GTP because it basically collapses if we give up the rich language assumption. Here are two examples which illustrate these problematic aspects of CMR.

With respect to the example in figure 2.15, Rabin already acknowledged that "CMR is disturbingly weak in some contexts where communication seems natural" (Rabin 1990, p. 158). CMR predicts that there will be no credible communication in this game, although intuitively we feel that t_3 would like to reveal himself while only t_1 and t_2 would surely pool together.[50] Unfortunately, CMR cannot predict this half-communication outcome, because there is no

48. Rabin makes such a rich language assumption, much in the spirit of Farrell (1993), but he also allows several messages $m \neq m'$ with $[\![m]\!] = [\![m']\!]$. To assume that there is a single unique message for each subset of states is just to simplify the exposition, but does not change anything of substance.

49. Rabin is particularly interested in sets of credible messages that induce a maximal subset and a maximally coarse-grained partition in this sense, because this gives a measure of the maximal amount of trustworthy, revealing communication that is possible in a given game. Rabin shows that such a set always exists.

50. I restrict attention here to the three messages given in figure 2.15. It does not change anything to keep the rich language assumption in place.

	a_1	a_2	a_3	a_4	m_1	m_2	m_3
t_1	7,6	6,7	0,0	-1,5	$\surd$	-	-
t_2	6,7	7,6	0,0	-1,5	-	$\surd$	-
t_3	0,0	0,0	6,6	-1,5	-	-	$\surd$

Figure 2.15: Illustration Rabin-credibility

Rabin-credible message. Not even m_3 is Rabin-credible, because $Y^*(m_3) = \{t_1, t_2\}$, in turn because

$$R_0(m_1) = \{a_1\} \neq \{a_4\} = \arg\min_{a \in A} \mathrm{U}_S(t_1, a)$$

and similarly for m_2. (The intuition in the background is that m_3 is Rabin-incredible, because it might be sent in t_1 or t_2 by a sender who wants to prevent her worst possible outcome.) But without any Rabin-credible message at all, CMR is equivalent to weak rationalizability. Effectively, CMR predicts that any strategy profile is a viable solution for this case.

On the other hand, IBR predicts that m_3 is credible and that partly revealing communication will ensue in this situation. Here is the R_0-sequence for illustration (the S_0-sequence brings no surprises either):

$$\begin{aligned}
R_0 &= \left\{ \begin{array}{lcl} m_1 & \mapsto & a_2 \\ m_2 & \mapsto & a_1 \\ m_3 & \mapsto & a_3 \end{array} \right\} \\
S_1 &= \left\{ \begin{array}{lcl} t_1 & \mapsto & m_2 \\ t_2 & \mapsto & m_1 \\ t_3 & \mapsto & m_3 \end{array} \right\} \\
R_2 &= \left\{ \begin{array}{lcl} m_1 & \mapsto & a_1 \\ m_2 & \mapsto & a_2 \\ m_3 & \mapsto & a_3 \end{array} \right\} \\
S_3 &= \left\{ \begin{array}{lcl} t_1 & \mapsto & m_1 \\ t_2 & \mapsto & m_2 \\ t_3 & \mapsto & m_3 \end{array} \right\} \\
R_4 &= R_0.
\end{aligned}$$

This example suggests that the requirements for Rabin-credibility are rather high, thus the resulting refinement of rationalizability sometimes too weak for reasonable predictions.

This problem also surfaces when we drop the rich language assumption and try to apply CMR to the simple some-all game for scalar implicature. Again, all available messages turn out Rabin-incredible and CMR consequently collapses into weak rationalizability, predicting a total anything-goes for sender and receiver behavior. To see that this is so, let us first check that $m_{\texttt{some}}$ is incredible. It is, because it fails the first condition of Rabin-credibility: intuitively, literal uptake does not induce *exactly* the set of sender best outcomes in all states where it is true. (It is thus obvious how CMR is designed towards a rich language assumption and how Rabin-credibility excludes pragmatic enrichment.) But also $m_{\texttt{all}}$ is Rabin-incredible, because it fails the second condition of the definition: in particular $Y^*(m_{\texttt{all}})$ does contain $t_{\exists\neg\forall}$ because it is not true that $m_{\texttt{some}}$ induces only *strictly* worse sender-outcomes under a literal uptake (intuitively this means that $m_{\texttt{all}}$ might be used also by type $t_{\exists\neg\forall}$ under a literal interpretation). Consequently, both messages are Rabin-incredible and hence CMR not only fails to select the scalar implicature play uniquely, but actually fails to restrict the strategy space entirely.

SUMMARY. The upshot of the above comparison of IBR, Farrell-, Stalnaker-credibility and CMR seems to be this: IBR implements basic ideas underlying all of these approaches in a formally efficient and intuitively accessible manner, while even improving on predictions about credibility and pragmatic inferences at the same time. The IBR model also compares well with rationalizability, to which it adds focality of semantic meaning and the assumption of unbiased belief formation, which is both a simplification of the formalism as well as an implementation of a forward induction rationale. Having thereby located and delineated the model conceptually, it remains to be shown that it also fares well in linguistic applications.

Chapter 3

Games and Pragmatic Interpretation

> "Informants, particularly those of the *ti'yč̵r* caste, claim this distinction to be present in their language; but when concrete evidence in support of this contention was demanded, informants invariably resorted to evasive and mystical references to context, sentence structure, collocational meaning, and the like." (Walker 1970, p. 104)

Chapter Contents

3.1 · Game Models Revisited · 124

3.2 · Epistemic Lifting of Signaling Games · 138

3.3 · Free Choice Inferences · 156

3.4 · Games at the Semantics-Pragmatics Interface · 175

This chapter deals with applications of the IBR model. Section 3.1 first discusses how to set up and interpret context models for generic interpretation of sentences. Section 3.2 then introduces the idea of epistemic lifting of signaling games that will help us integrate sender uncertainty into the model. Section 3.3 shows how such lifted models figure in an account of free choice readings. Finally, section 3.4 places this game theoretic approach into a broader linguistic context.

3.1 Game Models Revisited

The predictions of GTP not only depend on a suitable solution concept, but also on how the game model is set up. Parameter choices in a signaling game, such as the individuation of states, the set of messages, prior probabilities, utilities and so forth should therefore be based on uniform and generally motivated principles. This crucial aspect is unfortunately not addressed with due explicitness in the bulk of the relevant literature (except maybe Benz 2009) and so we should spent some thought on the issue.

3.1.1 Interpretation Games

Epistemic Status of the Context Model. Remember that with a game we want to capture the relevant aspects of the context of utterance. In game theory, the structure of a game is usually taken to be common knowledge between players. If taken at face value, this is certainly a dubious assumption for models of natural language interpretation (cf. Clark and Marshall 1981; Sperber and Wilson 1995; Allott 2006). It would be difficult to find many naturally occurring conversational settings where it is common knowledge between interlocutors that the context is as described by a signaling game.

But we also do not have to interpret the context model this way. We should rather consider the game model as a representation such that either sender or receiver —depending on whose side we wish to focus on— believes that it is common belief that the context is like that. This way the game represents the epistemic situation of only one agent separately and we make room for the possibility of subjective error. In what follows I will make the case that signaling game models in GTP should in particular be regarded as the *receiver's* conceptualization of the context *after* he has received a given target message whose interpretation we, as modellers, are interested in.

Interpretation Games. From this point of view, GTP is concerned with very particularized instances of individual reasoning about language use and interpretation. Still, GTP is not exclusively about situated reasoning in a particularized context. It can also be applied to a more abstract and general setting, and indeed this is what we have done implicitly already in the previous chapters. If we wish to speak of an "implicature of a given sentence" then we would like to resort to a more abstract notion of a *generic context model*. The idea is that, strictly speaking, sentences as such do not have implicatures, but that utterances of sentences in context have, and that to speak of "implicatures of sentences" is to be sloppy and to refer to implicatures that a sentences has in a standard, run-of-the-mill context (see Bach 1999).

The generic context model for natural language interpretation that I favor in this thesis is that of an interpretation game. An INTERPRETATION GAME is a signaling game where (i) the receiver's response actions are INTERPRETATION ACTIONS $A = T$ that correspond one to one with the set of states and where moreover (ii) response utilities are given as:

$$\mathrm{V}_S(t,a) = \mathrm{V}_R(t,a) = \begin{cases} 1 & \text{if } t = a \\ 0 & \text{otherwise.} \end{cases}$$

If message costs apply in interpretation games, these are *nominal* sending costs incurred equally for all states: $\mathrm{C}_S(t,m) = \mathrm{C}_S(t',m)$ (see section 2.2.4 on nominal costs). Since interpretation games aim to explain generic interpretation of sentences, these games are to be understood as representations of in particular the *receiver's* conceptualization of the utterance context.

For instance, the some-all game, which has accompanied us through the last two chapters already, should be thought of as an interpretation game for a generic utterance of a sentence like (8a). It is as such a representation of a (generic) receiver belief —be that a stipulation, an implicit assumption or actual well-grounded knowledge— that it is common belief between sender and receiver that the context of utterance is as given in the some-all game.

Interpretation of Interpretation Games. Interpretation games thereby implement a number of assumptions —of the interpreter if you wish— about the utterance context, most of which are familiar from other work in linguistic pragmatics. Firstly, interpretation games assume that the receiver is interested to get to know *precisely* which state is actual. This is implemented by interpretation actions $A = T$ and the receiver's utilities that match states and actions one-to-one. In general, we should think of the receiver's actions and

his payoffs as a fairly flexible way of representing a contextual *question under discussion* (see van Rooij 2003*b*). Receiver actions and payoffs fix what (kind of) information will be important and, indeed, *relevant* to the receiver.

Secondly, while the receiver's utilities implement the question under discussion, the sender is assumed to adopt the receiver's preferences in interpretation games. To assume that preferences of sender and receiver are aligned in the form of $V_S = V_R$ is a formal means of implementing Grice's central assumption that conversation is a (by and large) cooperative endeavour: in our game models, it is in particular the speaker who thus tends to the hearer's informational needs.

Thirdly, interpretation games assume that if messages are to be distinguished from one another by costs, then these costs are to be nominal. This is to implement the idea that whatever it is that makes one message more marked than another, first and foremost interlocutors strive to be understood. Only when they are otherwise indifferent would considerations of markedness of expressions play a role.[1]

It is of course not necessary that these assumptions all apply for all cases of natural language interpretation. But if we want to account for the implicatures associated with utterances of certain sentences in standard contexts, these assumptions do seem feasible.

3.1.2 Construction of Interpretation Games

By definition, interpretation games determine the receiver's response actions and the utilities of both sender and receiver. This still leaves other parameters of the context model underspecified. How should we interpret and specify the set of states, the prior probabilities, the set of messages, their meaning and possible message costs? I will argue in more detail below that an interpretation game for a given sentence should be constructed as follows: (i) a set of alternative forms to the given target sentence fixes the set of messages; (ii) the semantic meaning of messages is given by some suitable semantic theory; (iii) from the set of messages and their semantics, a set of state distinctions is derived; (iv) finally, prior probabilities over states are assumed flat. Let us have a closer look at these rough construction steps one by one in order to see what their respective motivations and conceptual implications are.

1. This assumption is not very crucial. Nothing hinges on this, as far as predictions of the IBR model are concerned for the examples considered here. Still, nominal message costs are (i) conceptually defensible and (ii) easier to cope with in proofs of structural results.

Message Alternatives. It is a well-known problem in Gricean pragmatics that naïve scalar reasoning (see section 1.1.2) depends heavily on the proper specification of alternative expressions $\text{Alt}(X)$. To wit, the inference from (8a) to (8g), repeated here, hinges on the idea that the speaker would have used the semantically stronger (8d) if it had been true and relevant.

(8a) Some of Kiki's friends are metalheads.

(8d) All of Kiki's friends are metalheads.

(8g) $\rightsquigarrow$ It's not the case that all of Kiki's friends are metalheads.

But by the same reasoning we could establish that (8a) implicates (24), because the speaker has not uttered the semantically stronger sentence (23) in which "some" is replaced by the expression "some but not all."

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

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

It is thus clear that naïve scalar reasoning crucially hinges on the set of alternatives that we take into account. This problem has been discussed centrally in Neo-Gricean pragmatics (Atlas and Levinson 1981; Horn 1989; Matsumoto 1995), and has recently been revived and dubbed symmetry problem (e.g. Fox 2007; Katzir 2007; Block 2008). Although for most cases there is usually a commonly shared understanding of which forms are a natural set of expression alternatives, it is fair to say that after all these years there is still no general consensus in the literature exactly *why* certain interfering expressions, like "some but not all" as an alternative to "some", should be excluded from naïve scalar reasoning. Suggestions why this may be so range from a different degree of lexicalizaion (Atlas and Levinson 1981), over a difference in monotonicity properties (Horn 1989), to increased structural complexity (Katzir 2007), for instance.

It is obvious that this issue is strictly speaking orthogonal to the concerns of gtp, which is a theory of *reasoning about* alternative messages and not a theory of alternatives as such. In order to account for an utterance of sentence X, the most natural choice for gtp is to stick —as uncommitted as possible— to the most natural and hopefully uncontroversial set $\text{Alt}(X)$ in order to derive its set M of speaker options. So, for instance, for a scalar implicature associated with the expression "some" we could assume that the set of alternatives are, in the vein of Horn (1989), all those semantically stronger sentences obtained from replacing "some" in the target utterance with a lexically, or otherwise, related expression. In sum, I would like to rely on a commonsense

notion of expression alternatives here, relegating the symmetry problem, as well as the question whether the present approach deals with *all* reasonable solutions to the symmetry problem, to another occasion.

States and Semantic Meaning. Based on a set of messages M we will also have to provide a semantic denotation function $[\![\cdot]\!]$, which we assume is a function from messages to sets of states. The most natural way of thinking about this is certainly that a state $t \in T$ of an interpretation game is a "state of the world," and that $t \in [\![m]\!]$ whenever the message m is true in t. It should be noted, however, that we do not have to interpret semantic meaning as truth conditions in this way. I will do so for all applications in this thesis, but strictly speaking, the signaling game model is compatible with different notions of semantic meaning, as long as the meaning of a message can be reasonably expressed in set theoretic terms, in particular as a subset of T.

For any arbitrary semantics that we would like to feed into our GTP model, the set of states of our interpretation game would then serve to represent all those distinctions in meaning that we would like to make based on the set M for the purposes of the current application. Under my preferred interpretation of conventional meaning as truth conditions, the set of states T should then be regarded as a set of disjoint sets of worlds, so as to represent all those relevant mutually exclusive states of affairs that can be expressed by the linguistic means provided by M.

More concretely, if we want to explain the scalar implicature associated with some (expression that we represent as) message m^*, then, by our conventional construction of the set M, it is usually the case that m^* is the semantically weakest message in M, in the sense that all other messages in M entail m^*. In that case it is natural to construe the set of states T as a partition of the set of worlds in which m^* is true so as to represent certain finer meaning distinctions as possible interpretations of the target expression m^*. But then, not every such partition is reasonable. Rather we should naturally consider exactly those distinctions that we can express with alternative messages beyond m^* under logical conjunction and negation. Effectively, we are interested in any meaning distinctions that be expressed by formulas of the kind:

$$m^* \wedge m_1 \wedge \ldots \wedge m_i \wedge m_{i+1} \wedge \ldots \wedge m_{i+j}$$
$$\vdots$$
$$m^* \wedge m_1 \wedge \ldots \wedge m_i \wedge \neg m_{i+1} \wedge \ldots \wedge \neg m_{i+j}$$
$$\vdots$$
$$m^* \wedge \neg m_1 \wedge \ldots \wedge \neg m_i \wedge \neg m_{i+1} \wedge \ldots \wedge \neg m_{i+j}$$

Some of these formulas will be inconsistent and some will be equivalent. Ultimately, we would then identify the states of our interpretation games with the non-contradictory propositions, i.e., non-empty sets of possible worlds, that can be expressed by a formula of the above list.

Prior Probabilities. After settling on a suitable set of state distinctions for our game model, we also need to specify a prior probability distribution on these. Remember that normally in game theory the prior probabilities in a signaling game would capture the receiver's initial beliefs about which state is actual. It is dubious, however, that this is a reasonable interpretation for applications to natural language interpretation, as I would like to argue in the following. I would also like to argue that the way prior probabilities have been used in previous approaches within GTP is similarly dubious. In conclusion, I suggest that we should look at prior probabilities as concise representations of the receiver's meaning associations *ex post*.

Let us begin by briefly revisiting the standard interpretation of prior probabilities in a signaling game. Game theorists like to think of the states of a signaling game as initial chance moves by a third player, called Nature, who selects any state $t \in T$ with probability $\Pr(t)$, without any strategic concern of her own (cf. Harsanyi 1967, 1968*a*,*b*). In a signaling game, Nature reveals her choice to only the sender, but not the receiver. According to this interpretation, the probability distribution $\Pr(\cdot)$ first and foremost captures the objective, frequentist probability of states actually occurring.

This standard interpretation of prior probabilities is not adequate for games as context models for natural language interpretation for two reasons: firstly, there are good arguments why objective, especially frequentist, probabilities are often not the driving force behind natural language disambiguation; secondly, it is moreover quite implausible for many linguistic applications that the receiver has any relevant probabilistic beliefs about states of affairs or

even intended speaker meanings *before* actually observing a message. Let me briefly enlarge on both of these points before offering an alternative view.

Consider the example (25), which is an example raised critically by Allott (2006) —borrowed from Wilson and Matsui (1998)— to argue against the use of prior probabilities in Parikh (2001)'s approach to GTP.

(25) John wrote a letter.

a. John wrote a letter of the alphabet.

b. John wrote a letter of correspondence.

The example demonstrates that we should not make the hearer's initial beliefs $\Pr(\cdot)$ responsible for the proper disambiguation of this sentence, as Parikh (2001) does.[2] In a normal context, (25b) is the preferred interpretation, and not (25a), whereas it is equally compelling to assume that in a normal context an unbiased hearer should hold it more likely that the proposition expressed in (25a) is true than that the proposition expressed in (25b) is.[3] So then if the proper disambiguation of (25) is to be achieved by modelling the context so that the state corresponding to (25b) has a higher prior probability, then obviously $\Pr(\cdot)$ should not express the receiver's initial belief which state of affairs is more likely, as derived —normatively correct— from objective chance. The example suggests that we should not conflate the receiver's possible conjectures about actualities with his preferred interpretations of expressions.

In response to this problem, we could assume, as Parikh (2001) suggests, that the prior probabilities should be taken to represent a conjecture about the sender's intended meaning. We could in fact take the states of the signaling game to be states that are individuated by the proposition which the speaker intends to express. This is fine for the conceptualization of states, but not necessarily for an interpretation of prior probabilities. Why would anyone have prior convictions about a speaker's communicative intentions irrespective of an utterance before anything has been said? It thus seems that the only

2. Similar conceptual issues arise for accounts of I-implicatures (see section 1.1.3) in terms of initial probabilities, as for instance done by Franke (2008*a*) or Jäger (2008*c*) in GTP, or by Blutner (1998) in bidirectional optimality theory. Similar concerns to the issues raised here were also discussed in the context of bidirectional optimality theory with respect to the *om/rond* problem for iconic interpretation (see Zwarts 2006).

3. Please bear with the assumptions that (i) John is not proficient in any non-alphabetic writing system, such as logographic or syllabic, and that (ii) John conforms to the common practice of writing letters of correspondence which consist of more than one alphabetic letter, whenever he writes one, or alternatively that some of his writings are not letters of correspondence.

straightforward appeal to probabilities is that in some contexts one interpretation of (25) is more likely than the other and in other contexts it may be the other way around. But this is not necessarily to be expressed in the receiver's initial beliefs, if these are explicit beliefs the receiver holds *before* a message is observed. It is also not something that we want to simply feed into the model as a contextual parameter, or else we should not conceive of this as an account of the disambiguation process.

It appears that in some contexts —think: out-of-the-blue utterances— the prior beliefs of the receiver could best be regarded as merely a condensed, systematic specification of the beliefs that the receiver holds *after* he receives an utterance. It would then be reasonable to draw on information in prior probabilities for models of disambiguation whenever it is feasible to assume that the receiver's posterior probabilistic beliefs effect the interpretation of ambiguous sentences like (25). But even this may at times be dubious: some cases of actual ambiguity are, introspectively speaking, *not* cases of PERCEIVED AMBIGUITY (cf. Poesio 1996). In other words, although the model suggests otherwise, the receiver might not have introspective access to the interpretation (25a) in a given standard context at all; we might want to say here that the receiver need not even be *aware* of the ambiguity.

This suggests that the prior probabilities in the game model might for some applications be understood rather as a measure for the probability with which certain interpretations *spring to mind* after a given target utterance has been observed. Interestingly, this interpretation of prior probabilities in our model matches a hypothesis of Tversky and Kahnemann (1983), who suggest that subjective probability, as measured under laboratory conditions, reflects the ease with which certain contingencies come to mind. For instance, when judging the subjective probability with which a woman can become CEO of an important company, we actually assess the relative ease with which examples of female CEOs come to mind, as compared to male CEOs. Similarly, the prior probabilities in a context model are a compact way of representing the relative ease with which interpretations associate with given messages. Indeed, to my mind, if prior probabilities are used in game models to differentiate between states, this seems like the most appealing conceptualization.

What does this interpretation of prior probabilities entail for explanations obtained from GTP, and for the way we should set up a game model for interpretation? Firstly, it transpires that this conceptual reinterpretation does not truly dispel our worry that disambiguation in terms of probabilities — however interpreted— is not much of an explanation at all: the disambigua-

tion remains a direct function of the parameter $\Pr(\cdot)$ which is fed in by hand. One possible and, to my mind, realistic conclusion here could be that GTP does not offer a full account of all processes involved in natural language interpretation, such as example (25), or standard cases of I-implicatures. It is for this reason that this thesis does not explicitly address the interpretation of I-implicatures in game theoretic terms.[4]

The upshot of this discussion is that it is contentious whether we should use information in prior probabilities at all as an explanatory element in game models for natural language *interpretation*. I therefore tend to believe that interpretation games, especially when used for explanations of scalar implicatures, should *not* contain any prior probability distinctions whatsoever. In fact, I propose to assume *flat priors* in interpretation games wherever possible.[5] Given the slack in interpretation of prior probabilities, this assumption first and foremost keeps our models simple and conceptually sober. Additionally, a flat prior assumption could also be appealed to based on empirical arguments: although the "principle of insufficient reason" may or may not be compelling from a normative point of view (see Keynes 1921, chapter 4), it is a good first shot at people's actual probabilistic belief formation in the absence of any further information that could influence a probabilistic judgement (cf. Falk 1992; Johnson-Laird et al. 1999; Fox and Levav 2004). So, especially for generic contexts for the interpretation of scalar expressions, I will assume flat priors. Under an interpretation of prior probabilities as strength of association between meanings and forms, the assumption of flat priors in (cheap talk) interpretation games comes down to the assumption that we do not rely on any *ex ante* associations between meanings and forms beyond (mostly: truth-

4. Nonetheless, GTP could still do *some* reasonable work in cases that hinge on various degrees of association of interpretations and expressions. Even if GTP as I conceive it may not explain very well the receiver's pragmatic inference in disambiguation of a sentence like (25) —perhaps, because there is no *inference* to begin with— the approach might still explain the speaker behavior: obviously, under my preferred interpretation of prior probabilities, it is rather the sender who has to reason about associations of possible utterances, than the receiver who has to reason about the likelihood of intended meanings; that means that we would predict the speaker to use less economic, more prolix formulations to increase the chance of correctly associating an utterance with the right interpretation, while in other contexts relying —successfully or not— on shorter, more economic but less specific formulations.

5. Notice that the assumption of flat priors interferes with the individuation of states: for instance, in the interpretation game for an utterance such as "Some of the (three) children are dirty", we could either assume two states, $t_{\exists\neg\forall}$ and $t_{\forall}$ as before, or assume that there are three states giving the number of dirty children. In the latter case the proposition that not all children are dirty has prior probability $^2/_3$ under a flat prior assumption.

conditional) semantic meaning, i.e., that all pragmatic enrichments merely stem from reasoning about the 'structural properties' of the triple M, T and $[\![\cdot]\!]$.

Game Construction Ex Post. In conclusion, I propose to think of an interpretation game as basically constructed *ex post* from the to-be-interpreted target expression X. A set of alternatives $\text{Alt}(X)$ will naturally suggest itself, or will otherwise be supplied by a theory of salience, lexical association or the like. From this set we construct a set of states that captures the meaning distinctions that we can express with the linguistic expressions available. Generic interpretation games should moreover ideally assume flat priors in the absence of reasons to model context-specific associative biases between meanings and expressions. As a whole, the game model should be considered the receiver's conceptualization of the context of utterance as triggered by observation of the target expression.

A number of remarks are in order concerning this model construction *ex post*. Firstly, we find that a given interpretation game, like the M-implicature game, is clearly *asymmetric*, in the sense that, for instance, the M-implicature game is a game for interpreting the long form $m_{\texttt{lng}}$, but that does not necessarily mean that we would arrive at the same game model for an interpretation of the short form $m_{\texttt{shrt}}$. This is entirely intuitive: when processing an unmarked expression, we do not necessarily reason about more complex ways of saying the same thing; in other words, the interpretation of the short form in the M-implicature game may be safely thought of as a fairly direct I-implicature without comparison to a more complex expression, while only the interpretation of the long form involves reasoning about alternatives.

Secondly, constructing interpretation games from a target expression also explains why certain features are missing in the representation of the context. For example, the some-all game does not contain a state $t_{\neg\exists}$ and also no form $m_{\texttt{none}}$. But that does not mean that it is impossible for this state to become actual, or more crucially even, that it is impossible for the sender to use $m_{\texttt{none}}$. If we were to take the some-all game at face value, it may seem that we imagine a context in which the speaker could only ever make a choice between saying $m_{\texttt{all}}$ or $m_{\texttt{some}}$. But in real life speakers make much more elaborate decisions, of course: they may continue, raise or zoom in on various topics, ask questions, or keep their mouths shut. All of these alternative actions are excluded from our generic game models, for which they are considered irrelevant, because such considerations do not play a role in the generic re-

ception of sentences. Still, at particular occasions, interpreters might ponder what a speaker has meant by comparing an utterance to the speaker's possibility of not speaking at all, or saying something entirely different (see also sections 5.2.4 and 5.3.4 on the interpretation of conditionals in context). This could easily be integrated into game models of particular utterance contexts, but in normal, generic interpretation cases this does not seem necessary. The crucial point is that particular game models for interpretation are different from generic game models for interpretations, and both may occasionally be different still from game models for production.

3.1.3 Examples: Multiple Scalar Items

Let us see how the principles of generic model construction apply to some relevant examples. In particular, let us look at examples that include multiple scalar items, once with independent scope, and once with nested scope. Examples of this kind have been addressed by Chierchia (2004) as critical for naïve scalar reasoning, but it will turn out that the IBR model deals with these cases effortlessly.

Independent Scope

Consider example (26a) with its intuitively attested implicatures in (26b)–(26d).

(26) a. Kai ate some of the strawberries and Hannes ate some of the carrots. (Sauerland 2004, ex. (24))

b. $\rightsquigarrow$ It's not the case that Kai ate all of the strawberries and Hannes ate some of the carrots.

c. $\rightsquigarrow$ It's not the case that Kai ate some of the strawberries and Hannes ate all of the carrots.

d. $\rightsquigarrow$ It's not the case that Kai ate all of the strawberries and Hannes ate all of the carrots.

For the target sentence (26a), which I will represent as $m_{\texttt{some|some}}$, there is not much controversy that the most natural set of alternative expressions and with it our set M should consist of (26a) itself together with the obvious scalar alternatives in (27).

(27) a. Kai ate some of the strawberries and Hannes ate all of the carrots. ($= m_{\texttt{some|all}}$)

b. Kai ate all of the strawberries and Hannes ate some of the carrots.
(= $m_{\texttt{all}|\texttt{some}}$)

c. Kai ate all of the strawberries and Hannes ate all of the carrots.
(= $m_{\texttt{all}|\texttt{all}}$)

If we assume the normal truth-conditional semantics and the canonical construction procedure outlined in the previous section, we want to consult all conjunctions that contain $m_{\texttt{some}|\texttt{some}}$ and all alternatives, either negated or not. The consistent formulas in this list will yield our state distinctions:

	$m_{\texttt{some}\|\texttt{all}}$	$m_{\texttt{all}\|\texttt{some}}$	$m_{\texttt{all}\|\texttt{all}}$
$t_{\forall\|\forall}$	$\surd$	$\surd$	$\surd$
incons.	$\surd$	$\surd$	$-$
incons.	$\surd$	$-$	$\surd$
$t_{\exists\neg\forall\|\forall}$	$\surd$	$-$	$-$
incons.	$-$	$\surd$	$\surd$
$t_{\forall\|\exists\neg\forall}$	$-$	$\surd$	$-$
incons.	$-$	$-$	$\surd$
$t_{\exists\neg\forall\|\exists\neg\forall}$	$-$	$-$	$-$

The table lists all possible conjunctions of $m_{\texttt{some}|\texttt{some}}$ with possibly negated stronger messages. Some of these constellations are inconsistent. Those that are not are given mnemonic state names. For example, state $t_{\exists\neg\forall|\forall}$ is the set of all worlds where the target message $m_{\texttt{some}|\texttt{some}}$ is true, where $m_{\texttt{some}|\texttt{all}}$ is also true and where $m_{\texttt{all}|\texttt{some}}$ and $m_{\texttt{all}|\texttt{all}}$ are false.

In the absence of any reason to assume differences in prior probabilities or in message costs we arrive at the interpretation game in figure 3.1. This game bears no surprises for the IBR model. As is easy to verify, the unique fixed point interpretation strategy predicts the intuitively attested implicatures:

$$R^* = \left\{ \begin{array}{lcl} m_{\texttt{some}|\texttt{some}} & \mapsto & t_{\exists\neg\forall|\exists\neg\forall} \\ m_{\texttt{some}|\texttt{all}} & \mapsto & t_{\exists\neg\forall|\forall} \\ m_{\texttt{all}|\texttt{some}} & \mapsto & t_{\forall|\exists\neg\forall} \\ m_{\texttt{all}|\texttt{all}} & \mapsto & t_{\forall|\forall} \end{array} \right\}.$$

Nested Scope

A slightly different and more interesting case is example (28a) where a scalar item occurs in the scope of another scalar item.

	$\Pr(t)$	$t_{\exists\neg\forall\|\exists\neg\forall}$	$t_{\exists\neg\forall\|\forall}$	$t_{\forall\|\exists\neg\forall}$	$t_{\forall\|\forall}$
$t_{\exists\neg\forall\|\exists\neg\forall}$	$1/4$	1,1	0,0	0,0	0,0
$t_{\exists\neg\forall\|\forall}$	$1/4$	0,0	1,1	0,0	0,0
$t_{\forall\|\exists\neg\forall}$	$1/4$	0,0	0,0	1,1	0,0
$t_{\forall\|\forall}$	$1/4$	0,0	0,0	0,0	1,1

	$m_{\texttt{some}\|\texttt{some}}$	$m_{\texttt{some}\|\texttt{all}}$	$m_{\texttt{all}\|\texttt{some}}$	$m_{\texttt{all}\|\texttt{all}}$
$t_{\exists\neg\forall\|\exists\neg\forall}$	$\surd$	–	–	–
$t_{\exists\neg\forall\|\forall}$	$\surd$	$\surd$	–	–
$t_{\forall\|\exists\neg\forall}$	$\surd$	–	$\surd$	–
$t_{\forall\|\forall}$	$\surd$	$\surd$	$\surd$	$\surd$

Figure 3.1: Interpretation game for example (26a)

(28) a. Some of the students read some of the books.

b. $\rightsquigarrow$ It's not the case that all of the students read some of the books.

c. $\rightsquigarrow$ It's not the case that some of the students read all of the books.

d. $\rightsquigarrow$ It's not the case that all of the students read all of the books.

This example is structurally equivalent to Sauerland's example (30), who predicts and defends the implicatures in (28b)–(28d). An example of this kind is also discussed as a problem for game theoretic accounts of Gricean communication by Rothschild (2008). What appears problematic for this example is that the most natural set of alternatives:

$$M = \left\{ m_{\texttt{some}|\texttt{some}}, m_{\texttt{some}|\texttt{all}}, m_{\texttt{all}|\texttt{some}}, m_{\texttt{all}|\texttt{all}} \right\}$$

carves up the logical space into a partition that contains *more* elements than messages. It is then unclear how a theory of rational language use will assign these extra meanings to the available messages.

First things first. Let us first properly establish the set of state distinctions from our canonical construction rule. It turns out that we get all the states that we also found in the previous case with independent scope, but that there is one additional state distinction $t_{\forall\exists\&\exists\forall}$ possible in this case:[6]

6. Notice that I write $\exists$ here for $\exists\neg\forall$ to maintain readability.

	$m_{\texttt{some}\mid\texttt{all}}$	$m_{\texttt{all}\mid\texttt{some}}$	$m_{\texttt{all}\mid\texttt{all}}$
$t_{\forall\forall}$	$\surd$	$\surd$	$\surd$
$t_{\forall\exists\&\exists\forall}$	$\surd$	$\surd$	$-$
incons.	$\surd$	$-$	$\surd$
$t_{\exists\forall}$	$\surd$	$-$	$-$
incons.	$-$	$\surd$	$\surd$
$t_{\forall\exists}$	$-$	$\surd$	$-$
incons.	$-$	$-$	$\surd$
$t_{\exists\exists}$	$-$	$-$	$-$

With these sets of states, and the usual assumptions, we then arrive at the context model in figure 3.2. For this interpretation game we predict the following interpretation for both strands of the IBR sequence:

μ^*	$t_{\exists\exists}$	$t_{\exists\forall}$	$t_{\forall\exists}$	$t_{\forall\exists\&\exists\forall}$	$t_{\forall\forall}$
$m_{\texttt{some}\mid\texttt{some}}$	$3/4$	0	0	$1/4$	0
$m_{\texttt{some}\mid\texttt{all}}$	0	$3/4$	0	$1/4$	0
$m_{\texttt{all}\mid\texttt{some}}$	0	0	$3/4$	$1/4$	0
$m_{\texttt{all}\mid\texttt{all}}$	0	0	0	0	1

$$R^* = \left\{ \begin{array}{lcl} m_{\texttt{some}\mid\texttt{some}} & \mapsto & t_{\exists\exists} \\ m_{\texttt{some}\mid\texttt{all}} & \mapsto & t_{\exists\forall} \\ m_{\texttt{all}\mid\texttt{some}} & \mapsto & t_{\forall\exists} \\ m_{\texttt{all}\mid\texttt{all}} & \mapsto & t_{\forall\forall} \end{array} \right\}$$

The receiver does not rule out $t_{\forall\exists\&\exists\forall}$ as a possible interpretation for any of the messages which are true in this state, but he prefers the more specialized scalar implicature interpretations that are also intuitively attested. The additional state distinction that appeared problematic for Rothschild (2008)'s approach is not a problem for the IBR model.

Still, we should pause here for a moment and ask what exactly it is that we would call the IBR model's prediction of the receiver's interpretation. After all, the receiver's posterior beliefs $\mu^*(\cdot|m_{\texttt{some}\mid\texttt{some}})$ in the fixed point do not exclude the possibility $t_{\forall\exists\&\exists\forall}$. But since this possibility is less likely than $t_{\exists\exists}$, this is the interpretation action that the receiver chooses. That means that in some sense the IBR model does predict the attested implicatures in (28b) and (28c), and in another sense it does not. Which notion of "predicted interpretation" should we adopt, here and in general? I suggest that we should opt

	$\Pr(t)$	$t_{\exists\exists}$	$t_{\exists\forall}$	$t_{\forall\exists}$	$t_{\forall\exists\&\exists\forall}$	$t_{\forall\forall}$
$t_{\exists\exists}$	$1/5$	1,1	0,0	0,0	0,0	0,0
$t_{\exists\forall}$	$1/5$	0,0	1,1	0,0	0,0	0,0
$t_{\forall\exists}$	$1/5$	0,0	0,0	1,1	0,0	0,0
$t_{\forall\exists\&\exists\forall}$	$1/5$	0,0	0,0	0,0	1,1	0,0
$t_{\forall\forall}$	$1/5$	0,0	0,0	0,0	0,0	1,1

	$m_{\texttt{some}\mid\texttt{some}}$	$m_{\texttt{some}\mid\texttt{all}}$	$m_{\texttt{all}\mid\texttt{some}}$	$m_{\texttt{all}\mid\texttt{all}}$
$t_{\exists\exists}$	$\surd$	$-$	$-$	$-$
$t_{\exists\forall}$	$\surd$	$\surd$	$-$	$-$
$t_{\forall\exists}$	$\surd$	$-$	$\surd$	$-$
$t_{\forall\exists\&\exists\forall}$	$\surd$	$\surd$	$\surd$	$-$
$t_{\forall\forall}$	$\surd$	$\surd$	$\surd$	$\surd$

Figure 3.2: Interpretation game for example (28a)

for the receiver's preferred interpretation, which shows in his choice of interpretation action. This will yield the intuitively correct predictions not only for this example, but especially in later examples where probabilistic information on states is used to model different levels of speaker expertise in models that accommodate speaker uncertainty. This is the topic that we will turn to next.

3.2 Epistemic Lifting of Signaling Games

3.2.1 The Epistemic Status of Scalar Implicatures

Where previous chapters have dealt with scalar implicatures, the treatment so far has in fact been unduly simplistic in saying that an utterance of a sentence like (29a), when an alternative sentence (29b) could have been used, conveys the implicature in (29c).

(29) a. *Assertion:* Some of Kiki's friends are metalheads.

b. *Alternative:* All of Kiki's friends are metalheads.

c. *Factual Implicature:* It's not the case that *all* of Kiki's friends are metalheads.

d. *Weak Epistemic Implicature:* The speaker does not know/believe that all of Kiki's friends are metalheads. $\neg K_S(29b)/\neg B_S(29b)$

e. *Strong Epistemic Implicature:* The speaker knows/believes that it's not the case that all of Kiki's friends are metalheads.
$K_S(\neg(29b))/B_S(\neg(29b))$

Indeed, the actual logic of Gricean reasoning does not immediately give rise to the FACTUAL IMPLICATURE in (29c), but strictly speaking only allows the WEAK EPISTEMIC IMPLICATURE in (29d) to be drawn (see Gazdar 1979; Soames 1982): from an utterance of (29a) we can conclude that the speaker, though cooperative and willing to give all relevant information, was not in a position to utter the stronger sentence (29b); but that only licenses the inference that the speaker did not know or believe that (29b) was true. This inference is also often called the *primary implicature*, and it has become customary in the literature to assume that the stronger *secondary implicature* in (29e), which is also called the STRONG EPISTEMIC IMPLICATURE, is to be derived in a second step of reasoning from the weak epistemic implicature by the additional assumption that the speaker is competent or opinionated on the issue at hand (see Sauerland 2004; van Rooij and Schulz 2004; Schulz and van Rooij 2006; Russell 2006; Spector 2006). For the present example, to assume that the speaker is competent or opinionated in the relevant sense is to assume that she knows the truth value of sentence (29b) or at least has a decided, possibly prejudiced, belief about it:

(30) a. *Competence Assumption:* $K_S(29b) \vee K_S(\neg(29b))$
b. *Opinionated-Speaker Assumption:* $B_S(29b) \vee B_S(\neg(29b))$

The weak epistemic implicature in (29d) and the competence assumption (30) together establish the strong epistemic implicature in (29e). The factual implicature in (29c) may then be derived from the factivity of knowledge.

In previous chapters, we have ignored the particular epistemic status of the scalar inference in the IBR model. The analysis of scalar implicature so far accounts for the strong epistemic implicature and the factual implicature; and it does so immediately, i.e., without strengthening of the weak epistemic implicature by competence. This is so because a particular sort of sender competence assumption is already integrated in the basic signaling game model of the utterance context. Remember that in a signaling game (it is common knowledge between sender and receiver that) the sender knows the true state of affairs. Therefore, the implicature that we derive using standard signaling games must be the strong epistemic implicature and also, because the speaker's knowledge about the state of affairs is correct, the factual implicature.

But if standard signaling games incorporate sender competence as an assumption about the utterance context, that also means that there is no direct way of accounting for the weak epistemic implicature in (29d). This is indeed a shortcoming of the model as it stands, and it is, in particular, a shortcoming of our context models which are not (yet) flexible enough to accommodate the assumption that the speaker may only have *partial knowledge* about the relevant states of affairs. I will show presently that this shortcoming in the representation of utterance contexts obstructs proper predictions for other relevant kinds of scalar reasoning, in particular for the implicatures associated with standard uses of disjunction. Subsequently, I will present what is basically a conceptual reinterpretation of the signaling game framework which will allow us to keep the basic model as is, but nonetheless integrate a notion of speaker uncertainty into signaling games.

3.2.2 The Implicatures of Plain Disjunctions

The perhaps most prominent use of a disjunction is to express —in entirely intuitive terms— a list of epistemic possibilities not all of which are false (according to the speaker).[7] For instance, an utterance of the sentence in (31a) as an answer to the possibly implicit question "Who baked this cake?" seems to give rise to two kinds of inferences: the ignorance implicature in (31b) and the scalar implicature in (31c).[8]

(31) a. *Assertion:* This cake was baked by John or Mary. $A \vee B$

b. *Ignorance Implicature:* The speaker doesn't know whether John baked this cake and the speaker doesn't know whether Mary baked this cake. $\neg K_S A \wedge \neg K_S \neg A \wedge \neg K_S B \wedge \neg K_S \neg B$

c. *(Epistemic) Scalar Implicature:* The Speaker knows that it's not the case that John and Mary baked the cake together. $K_S \neg (A \wedge B)$

We would certainly like to derive both of these inferences also in the IBR model, but it is rather obvious that ignorance implicatures such as (31b)

7. There are other, less stereotypical uses of disjunctions, of course (see Culicover and Jackendoff 1997; Gómez-Txurruka 2002; Franke 2008*b*).

8. Whether the scalar implicature in (31c) usually arises or whether it requires a stressed *or* is a controversial matter. I will speak as if it ought to be derived generally. Nonetheless, the IBR model can also model the absence of this inference. Unsurprisingly, whether we derive this inference or not simply depends on whether we assume a message $m_{\mathtt{A}\wedge\mathtt{B}}$ in the signaling game model.

	$\Pr(t)$	$a_{\mathtt{A}}$	$a_{\mathtt{B}}$	$a_{\mathtt{AB}}$	$m_{\mathtt{A}}$	$m_{\mathtt{B}}$	$m_{\mathtt{A}\wedge\mathtt{B}}$	$m_{\mathtt{A}\vee\mathtt{B}}$
$t_{\mathtt{A}}$	$1/3$	1,1	0,0	0,0	$\surd$	$-$	$-$	$\surd$
$t_{\mathtt{B}}$	$1/3$	0,0	1,1	0,0	$-$	$\surd$	$-$	$\surd$
$t_{\mathtt{AB}}$	$1/3$	0,0	0,0	1,1	$\surd$	$\surd$	$\surd$	$\surd$

Figure 3.3: Unlifted interpretation game for example (31a)

are going to be problematic, as long as the speaker is assumed to be competent in the strong sense that she knows the true state of affairs. To make this point entirely clear, I suggest looking at the context model in figure 3.3.

For this game both IBR sequences terminate in the same fixed point in which the critical message $m_{\mathtt{A}\vee\mathtt{B}}$ is left dangling as a surprise message that is not used by the speaker:

$$S^* = \left\{ \begin{array}{lcl} t_{\mathtt{A}} & \mapsto & m_{\mathtt{A}} \\ t_{\mathtt{B}} & \mapsto & m_{\mathtt{B}} \\ t_{\mathtt{AB}} & \mapsto & m_{\mathtt{A}\wedge\mathtt{B}} \end{array} \right\} \qquad R^* = \left\{ \begin{array}{lcl} m_{\mathtt{A}} & \mapsto & t_{\mathtt{A}} \\ m_{\mathtt{B}} & \mapsto & t_{\mathtt{B}} \\ m_{\mathtt{A}\wedge\mathtt{B}} & \mapsto & t_{\mathtt{A}\wedge\mathtt{B}} \\ m_{\mathtt{A}\vee\mathtt{B}} & \mapsto & t_{\mathtt{A}}, t_{\mathtt{B}}, t_{\mathtt{AB}} \end{array} \right\}.$$

Even forward induction reasoning does not help, because the intuitive criterion does not rule out any state from the interpretation of $m_{\mathtt{A}\vee\mathtt{B}}$, because the sender would not gain anything in any state from sending this message compared to what she can obtain with the sending strategy S^*.

Hence, neither the ignorance implicature (31b), nor the scalar implicature (31c) can be accounted for in this model. The problem is clearly that with a perfectly informed speaker, i.e., a speaker who knows the truth values of propositions A and B, there is no room to derive implicatures relating to the speaker's epistemic uncertainty. This then calls for some amendment or extension of the standard model.

3.2.3 Lifted Signaling Games

The problem the IBR model is faced with is that (i) we would like to be able to derive *epistemic implicatures*, i.e., inferences that concern the speaker's epistemic uncertainty but that (ii) in a standard signaling game —or rather in the standard interpretation of a signaling game— this seems impossible to achieve, because the speaker is assumed fully knowledgeable about the true relevant state — an assumption which is a part of the game structure. So it may seem that we have to turn away from signaling games in order to include

speaker uncertainty as well. But this, it turns out, is not absolutely necessary. We can basically leave the model as it is but simply give an epistemic interpretation to states; in other words, we can *lift* the interpretation of states from states of the world to information states of the sender. Here is how it works:[9,10]

Information States. To account for the ignorance implicatures associated with an utterance of disjunction $A \vee B$ in (31a) we would assume that the set of states T is a set of information states. If

$$T_{\texttt{plain}} = \{t_{\texttt{A}}, t_{\texttt{B}}, t_{\texttt{AB}}\}$$

are the states of the plain, unlifted signaling game for disjunction in figure 3.3, then the set of all non-trivial information states is given as

$$T_{\texttt{lifted}} = \mathcal{P}(T_{\texttt{plain}}) \setminus \emptyset,$$

the set of all non-empty subsets of $T_{\texttt{plain}}$ — non-empty, because we should exclude the *absurd information state* in which the speaker does not consider *any* of the plain states possible, of course. I will then write $t_{[\texttt{A,AB}]} \in T_{\texttt{lifted}}$ for an information state in which the sender considers it possible that only $t_{\texttt{A}}$ or $t_{\texttt{AB}}$ might be the true (unlifted) states of affairs. In other words, $t_{[\texttt{A,AB}]}$ is alternative notation for $\{t_{\texttt{A}}, t_{\texttt{AB}}\}$.

Although the lifted states may represent the speaker's potential uncertainty, it is still feasible to maintain the assumption of standard signaling games that the sender (but not the receiver) knows which *lifted* state of affairs is actual: the sender simply knows which epistemic state she is in, even if that is a state of epistemic uncertainty about the state of the world. It is in this sense that by lifting the notion of a state to a representation of the sender's epistemic state we can integrate reasoning about the sender's partial knowledge into standard signaling games.

We then construct a lifted signaling game from the plain signaling game in figure 3.3 as follows. In the lifted game the set of messages remains the same. However, the interpretation of the semantic denotation function needs to be

9. Lifted signaling game models have also been employed by de Jager and van Rooij (2007) to rationalize the interpretation principle "Grice" from van Rooij and Schulz (2004). In general, the idea to look at the speaker's information states in pragmatic interpretation is of course very familiar (e.g. Gazdar 1979).

10. It needs to be stressed for clarity that the general lifting mechanism given here is strictly geared towards, and therefore possibly only makes sense for, interpretation games.

	$\Pr(t)$	$a_{[\mathtt{A}]}$	$a_{[\mathtt{B}]}$	$a_{[\mathtt{AB}]}$	$a_{[\mathtt{A,AB}]}$	$a_{[\mathtt{B,AB}]}$	$a_{[\mathtt{A,B}]}$	$a_{[\mathtt{A,B,AB}]}$
$t_{[\mathtt{A}]}$	a	1,1	0,0	0,0	0,0	0,0	0,0	0,0
$t_{[\mathtt{B}]}$	a	0,0	1,1	0,0	0,0	0,0	0,0	0,0
$t_{[\mathtt{AB}]}$	a	0,0	0,0	1,1	0,0	0,0	0,0	0,0
$t_{[\mathtt{A,AB}]}$	b	0,0	0,0	0,0	1,1	0,0	0,0	0,0
$t_{[\mathtt{B,AB}]}$	b	0,0	0,0	0,0	0,0	1,1	0,0	0,0
$t_{[\mathtt{A,B}]}$	b	0,0	0,0	0,0	0,0	0,0	1,1	0,0
$t_{[\mathtt{A,B,AB}]}$	c	0,0	0,0	0,0	0,0	0,0	0,0	1,1

	$m_{\mathtt{A}}$	$m_{\mathtt{B}}$	$m_{\mathtt{A}\wedge\mathtt{B}}$	$m_{\mathtt{A}\vee\mathtt{B}}$
$t_{[\mathtt{A}]}$	$\surd$	–	–	$\surd$
$t_{[\mathtt{B}]}$	–	$\surd$	–	$\surd$
$t_{[\mathtt{AB}]}$	$\surd$	$\surd$	$\surd$	$\surd$
$t_{[\mathtt{A,AB}]}$	$\surd$	–	–	$\surd$
$t_{[\mathtt{B,AB}]}$	–	$\surd$	–	$\surd$
$t_{[\mathtt{A,B}]}$	–	–	–	$\surd$
$t_{[\mathtt{A,B,AB}]}$	–	–	–	$\surd$

Figure 3.4: Lifted interpretation game for example (31a)

amended to accommodate the change in the notion of a state: in the lifted game, $[\![m]\!]$ yields the set of all information states where m is *believed true*. A message m is believed true in an information state t —considered as a set of non-lifted states of the world— if m is true in all non-lifted states contained in t. We should still treat the lifted game as an interpretation game, i.e., lifted states correspond to lifted interpretation actions and response utilities are given as $\mathrm{V}_{S,R}(t, m, a) = 1$ if $t = a$, otherwise 0. This is of course to represent the hearer's concern to understand the speaker's epistemic state expressed by an utterance of a sentence such as (31a). Taken together, the epistemically lifted signaling game derived is the one in figure 3.4.

Implementing Competence There is still one feature of the lifted signaling game model left unspecified: the receiver's prior beliefs about which epistemic state the speaker is most likely to be in. This is, I suggest, the natural place to implement a speaker competence assumption in the signaling game model similar to the one in (30), which we will need, just as previous Neo-Gricean approaches did, in order to tell, for instance, the weak and strong epistemic implicatures in (29d) and (29e) apart. I suggest that for epistemi-

cally lifted signaling games the competence assumption should take the following simple form:[11]

(32) *Competence Assumption:* If the speaker is (believed) competent, then smaller information states are strictly more likely than larger ones.

In other terms, if (the receiver believes that) the sender is competent, (the receiver believes that) the sender is more likely to rule out more alternatives. This is then reflected in the probabilities of lifted states and, therefore, the receiver's prior beliefs.[12]

However, the competence assumption in (32) still leaves some slack in the specification of prior probabilities. I will therefore stick to the previous considerations on flat priors in interpretation games additionally. In effect, where the sender competence assumption does not specify a difference in prior probabilities, probabilities are to be assumed equal. To be more precise, for a set of lifted states T whose elements are information states which we may consider sets of possible states of affairs, we assume that for all $t, t' \in T$:[13]

(i) if the speaker is competent (alt.: an expert), then

  (a) $\Pr(t) > \Pr(t')$ iff $|t| > |t'|$ and
  (b) $\Pr(t) = \Pr(t')$ iff $|t| = |t'|$;

(ii) if the speaker is not competent, then $\Pr(t) = \Pr(t')$.

In the example in figure 3.4, we should thus parametrize and distinguish cases of relative expertise. Let's set:

$$\Pr(t) = \begin{cases} a & \text{if } t \text{ contains 1 element} \\ b & \text{if } t \text{ contains 2 elements} \\ c & \text{if } t \text{ contains 3 elements.} \end{cases}$$

The speaker then is competent, just in case $a > b > c$; she is not competent for parameters $a = b = c$.

11. Alternatively, we could also define competence as a partial order on information states in terms of set inclusion, if that seemed more natural to us. However, mapping this ordering back onto (linear) probabilities in the most natural way will result in the exact same constraint on the receiver's prior as the notion I suggest here.

12. This does not interfere with my preferred interpretation of prior probabilities of *unlifted* states.

13. As long as T is finite, the question whether there exists any $\Pr \in \Delta(T)$ that satisfies these constraints has a trivial positive answer.

Notice that the pragmatic community is usually not concerned with assessing and systematizing intuitions about the contextual meaning of utterances under the assumption that the speaker is known to be an *in*expert: most of the time, the debate about (scalar) implicatures —implicitly or explicitly— assumes expert speakers; some of the time our intuitions are questioned about readings we obtain when we do not assume that the speaker is an expert; but, as far as I am aware, there is hardly any systematic investigation into the question what we would infer from the assumption that the speaker is *not* an expert. When discussing predictions of the IBR model I will also only explicitly discuss the former two cases, and I will even be sloppy in my choice of expression: a 'non-expert' or an 'incompetent speaker' is a speaker who is not assumed to be competent.[14]

3.2.4 Examples

Disjunction $A \vee B$

With these assumptions in place the IBR model yields a unique solution for the expert and a different unique solution for the non-expert case of the signaling game in figure 3.4. For experts the fixed point of *both* strands of the IBR model is given by:

$$S^* = \left\{ \begin{array}{lcl} t_{[\mathtt{A}]} & \mapsto & m_{\mathtt{A}} \\ t_{[\mathtt{B}]} & \mapsto & m_{\mathtt{B}} \\ t_{[\mathtt{AB}]} & \mapsto & m_{\mathtt{A}\wedge\mathtt{B}} \\ t_{[\mathtt{A},\mathtt{AB}]} & \mapsto & m_{\mathtt{A}}, m_{\mathtt{A}\vee\mathtt{B}} \\ t_{[\mathtt{B},\mathtt{AB}]} & \mapsto & m_{\mathtt{B}}, m_{\mathtt{A}\vee\mathtt{B}} \\ t_{[\mathtt{A},\mathtt{B}]} & \mapsto & m_{\mathtt{A}\vee\mathtt{B}} \\ t_{[\mathtt{A},\mathtt{B},\mathtt{AB}]} & \mapsto & m_{\mathtt{A}\vee\mathtt{B}} \end{array} \right\} \qquad R^* = \left\{ \begin{array}{lcl} m_{\mathtt{A}} & \mapsto & t_{[\mathtt{A}]} \\ m_{\mathtt{B}} & \mapsto & t_{[\mathtt{B}]} \\ m_{\mathtt{A}\wedge\mathtt{B}} & \mapsto & t_{[\mathtt{AB}]} \\ m_{\mathtt{A}\vee\mathtt{B}} & \mapsto & t_{[\mathtt{A},\mathtt{B}]} \end{array} \right\}$$

Let me just briefly spell out, for illustration, the R_0-sequence under a competence assumption. It is easy to verify that the numeric choice for prior probabilities does not heavily affect the qualitative outcome. For $a = 3/16$, $b = 1/8$ and $c = 1/16$ as parameters of the game in figure 3.4 we get:

14. Still, the IBR model in principle extends to and predicts readings also under an inexpert assumption. If we set $\Pr(t) > \Pr(t')$ iff $|t| < |t'|$, we derive for example that the form $m_{\mathtt{some}}$ in the some-all game gets to be interpreted as $t_{[\exists\neg\forall,\forall]}$ (for notation, see below). This seems a fine prediction for this simple case. Whether the IBR model predicts intuitively for other cases under such a strong inexpertise assumption is something I will have to leave for another occasion, especially also because I am not sure what the intuitively correct predictions should be in many of these cases.

μ_0	$t_{[\mathtt{A}]}$	$t_{[\mathtt{B}]}$	$t_{[\mathtt{AB}]}$	$t_{[\mathtt{A,AB}]}$	$t_{[\mathtt{B,AB}]}$	$t_{[\mathtt{A,B}]}$	$t_{[\mathtt{A,B,AB}]}$
$m_{\mathtt{A}}$	$3/8$	0	$3/8$	$1/4$	0	0	0
$m_{\mathtt{B}}$	0	$3/8$	$3/8$	0	$1/4$	0	0
$m_{\mathtt{A}\wedge\mathtt{B}}$	0	0	1	0	0	0	0
$m_{\mathtt{A}\vee\mathtt{B}}$	$3/16$	$3/16$	$3/16$	$1/8$	$1/8$	$1/8$	$1/16$

$$R_0 = \left\{ \begin{array}{lcl} m_{\mathtt{A}} & \mapsto & t_{[\mathtt{A}]}, t_{[\mathtt{AB}]} \\ m_{\mathtt{B}} & \mapsto & t_{[\mathtt{B}]}, t_{[\mathtt{AB}]} \\ m_{\mathtt{A}\wedge\mathtt{B}} & \mapsto & t_{[\mathtt{AB}]} \\ m_{\mathtt{A}\vee\mathtt{B}} & \mapsto & t_{[\mathtt{A}]}, t_{[\mathtt{B}]}, t_{[\mathtt{AB}]} \end{array} \right\}$$

for our naïve receiver. The competence assumption draws the receiver towards choosing the 'smallest' epistemic states compatible with each message's semantic meaning. Notice that S_1 can then not expect to induce the proper interpretation in any state where she is truly uncertain. By TCP assumption she will then be indifferent between all messages that are true in these states, so that we get:

$$S_1 = \left\{ \begin{array}{lcl} t_{[\mathtt{A}]} & \mapsto & m_{\mathtt{A}} \\ t_{[\mathtt{B}]} & \mapsto & m_{\mathtt{B}} \\ t_{[\mathtt{AB}]} & \mapsto & m_{\mathtt{A}\wedge\mathtt{B}} \\ t_{[\mathtt{A,AB}]} & \mapsto & m_{\mathtt{A}}, m_{\mathtt{A}\vee\mathtt{B}} \\ t_{[\mathtt{B,AB}]} & \mapsto & m_{\mathtt{B}}, m_{\mathtt{A}\vee\mathtt{B}} \\ t_{[\mathtt{A,B}]} & \mapsto & m_{\mathtt{A}\vee\mathtt{B}} \\ t_{[\mathtt{A,B,AB}]} & \mapsto & m_{\mathtt{A}\vee\mathtt{B}} \end{array} \right\}$$

The following best response of R_2 is again straightforward:

μ_2	$t_{[\mathtt{A}]}$	$t_{[\mathtt{B}]}$	$t_{[\mathtt{AB}]}$	$t_{[\mathtt{A,AB}]}$	$t_{[\mathtt{B,AB}]}$	$t_{[\mathtt{A,B}]}$	$t_{[\mathtt{A,B,AB}]}$
$m_{\mathtt{A}}$	$3/4$	0	0	$1/4$	0	0	0
$m_{\mathtt{B}}$	0	$3/4$	0	0	$1/4$	0	0
$m_{\mathtt{A}\wedge\mathtt{B}}$	0	0	1	0	0	0	0
$m_{\mathtt{A}\vee\mathtt{B}}$	0	0	0	$1/5$	$1/5$	$2/5$	$1/5$

$$R_2 = \left\{ \begin{array}{lcl} m_{\mathtt{A}} & \mapsto & t_{[\mathtt{A}]} \\ m_{\mathtt{B}} & \mapsto & t_{[\mathtt{B}]} \\ m_{\mathtt{A}\wedge\mathtt{B}} & \mapsto & t_{[\mathtt{AB}]} \\ m_{\mathtt{A}\vee\mathtt{B}} & \mapsto & t_{[\mathtt{A,B}]} \end{array} \right\}$$

To this the sender best responds with the above strategy $S_1 = S^*$ and a fixed point is reached.

Turning to the non-expert case, here both strands of the IBR model reach a fixed point in the following strategies:

$$S^* = \left\{ \begin{array}{lcl} t_{[\mathtt{A}]} & \mapsto & m_{\mathtt{A}} \\ t_{[\mathtt{B}]} & \mapsto & m_{\mathtt{B}} \\ t_{[\mathtt{AB}]} & \mapsto & m_{\mathtt{A}\wedge\mathtt{B}} \\ t_{[\mathtt{A,AB}]} & \mapsto & m_{\mathtt{A}} \\ t_{[\mathtt{B,AB}]} & \mapsto & m_{\mathtt{B}} \\ t_{[\mathtt{A,B}]} & \mapsto & m_{\mathtt{A}\vee\mathtt{B}} \\ t_{[\mathtt{A,B,AB}]} & \mapsto & m_{\mathtt{A}\vee\mathtt{B}} \end{array} \right\} \qquad R^* = \left\{ \begin{array}{lcl} m_{\mathtt{A}} & \mapsto & t_{[\mathtt{A}]}, t_{[\mathtt{A,AB}]} \\ m_{\mathtt{B}} & \mapsto & t_{[\mathtt{B}]}, t_{[\mathtt{B,AB}]} \\ m_{\mathtt{A}\wedge\mathtt{B}} & \mapsto & t_{[\mathtt{AB}]} \\ m_{\mathtt{A}\vee\mathtt{B}} & \mapsto & t_{[\mathtt{A,B}]}, t_{[\mathtt{A,B,AB}]} \end{array} \right\}$$

The calculation of this result is straightforward and I will refrain from giving details.[15] Suffice it to say that the ignorance implicature in (31b) that the speaker does not know of either disjunct that it is true is predicted for both experts and non-experts. However, the strong epistemic version of the scalar implicature in (31c) that the speaker knows that the disjuncts are not both true is derived only for the expert case; for non-experts we only derive a weak epistemic implicature that the speaker does not know whether both disjuncts are true at the same time.

I find these predictions intuitive and they also match the implicatures generally associated with standard informative uses of disjunctions in the literature. It should be stressed also that we did not have to assume any increased costs for the disjunctive form to obtain this result.

Scalar Implicatures

This pattern of explanation reoccurs in a straightforward lifting of the scalar implicature example (29a). We would like to account for both the weak and the strong epistemic implicature in (29d) and (29e) respectively. Lifting the standard signaling game for scalar implicature given in figure 1.3 in section 1.2.2, we obtain the signaling game model in figure 3.5. The notation for lifted states is as before: $t_{[\exists\neg\forall,\forall]}$ is an information state representing speaker uncertainty comprising the unlifted states $t_{\exists\neg\forall}$ and $t_{\forall}$. For experts we assume $a > b$ and for non-experts we assume $a = b$.

15. Results reported here are backed up by a computer implementation of the IBR model, which was obtained by amending code kindly provided by Gerhard Jäger who has been implementing his own version of IBR reasoning. In the text, I will preferably only give the 'analytical highlights' the first time they arise.

	$\Pr(t)$	$a_{[\exists\neg\forall]}$	$a_{[\forall]}$	$a_{[\exists\neg\forall,\forall]}$	$m_{\texttt{some}}$	$m_{\texttt{all}}$
$t_{[\exists\neg\forall]}$	a	1,1	0,0	0,0	$\surd$	–
$t_{[\forall]}$	a	0,0	1,1	0,0	$\surd$	$\surd$
$t_{[\exists\neg\forall,\forall]}$	b	0,0	0,0	1,1	$\surd$	–

Figure 3.5: Lifted scalar implicature model

Again the IBR model predicts a unique solution for both the expert and non-expert cases. The interested reader will find it easy to verify for herself that for expert senders the fixed point strategy pair is:

$$S^* = \left\{ \begin{array}{lcl} t_{[\exists\neg\forall]} & \mapsto & m_{\texttt{some}} \\ t_{[\forall]} & \mapsto & m_{\texttt{all}} \\ t_{[\exists\neg\forall,\forall]} & \mapsto & m_{\texttt{some}} \end{array} \right\} \qquad R^* = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_{[\exists\neg\forall]} \\ m_{\texttt{all}} & \mapsto & t_{[\forall]} \end{array} \right\}.$$

For non-experts we obtain:

$$S^* = \left\{ \begin{array}{lcl} t_{[\exists\neg\forall]} & \mapsto & m_{\texttt{some}} \\ t_{[\forall]} & \mapsto & m_{\texttt{all}} \\ t_{[\exists\neg\forall,\forall]} & \mapsto & m_{\texttt{some}} \end{array} \right\} \qquad R^* = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_{[\exists\neg\forall]}, t_{[\exists\neg\forall,\forall]} \\ m_{\texttt{all}} & \mapsto & t_{[\forall]} \end{array} \right\}.$$

This is the intuitively correct prediction.

McCawley-Chierchia Problem

Let us briefly venture into a less trivial example, which the IBR model solves effortlessly with the notion of epistemic lifting. The example in (33a) has been dubbed *Chierchia's puzzle* by Fox (2007) to credit the observation by Chierchia (2004) that an example like (33a) with scalar "some" in one of the disjuncts of a disjunction gives rise to a problem for naïve scalar reasoning, but a structurally parallel case has already been raised by McCawley (1981).

(33) a. Kai had the broccoli or some of the peas. $m_{\texttt{B}\vee\texttt{some}}$

b. *Ignorance Implicature: S* doesn't know whether Kai had broccoli.

c. *Ignorance Implicature: S* doesn't know whether Kai had some of the peas.

d. *Scalar Implicature 1: S* knows that Kai didn't have both broccoli and some of the peas.

e. *Scalar Implicature 2: S* knows that Kai didn't have all of the peas.

f. *Unattested Scalar Implicature: S* knows that Kai didn't have the broccoli.

Intuitively, sentence (33a) implicates (33b)–(33e), but *not* (33f), contrary to the predictions of naïve scalar reasoning. The misprediction arises, because the sentence (34) *is* a stronger alternative to the target sentence in (33a), but a negation of (34) actually entails the unattested inference in (33f).

(34) Kai had the broccoli or all of the peas. $m_{\mathtt{B}\vee\mathtt{all}}$

This example has been important in the recent literature on Gricean pragmatics, and this is why we should address it and see whether we get it right. The example was used as an argument against Neo-Gricean accounts of implicatures by Chierchia (2004) in order to support his view of local implicature calculation in the syntax. Responding to Chierchia's challenge, Sauerland (2004) subsequently defended a Neo-Gricean account.

The IBR model solves the McCawley-Chierchia problem without any complications. The correct predictions can be derived immediately after lifting a properly set-up context model. Next to the target sentence (33a) and the problematic alternative in (34), we should clearly include the messages in (35) as the set of alternative messages.

(35) a. Kai had the broccoli and some of the peas. $m_{\mathtt{B}\wedge\mathtt{some}}$

b. Kai had the broccoli and all of the peas. $m_{\mathtt{B}\wedge\mathtt{all}}$

c. Kai had the broccoli or all of the peas. $m_{\mathtt{B}\vee\mathtt{all}}$

d. Kai had the broccoli. $m_{\mathtt{B}}$

e. Kai had some of the peas. $m_{\mathtt{some}}$

f. Kai had all of the peas. $m_{\mathtt{all}}$

This set of alternatives gives rise to the set of states

$$T = \{t_{\mathtt{B}\exists\neg\forall}, t_{\mathtt{B}\forall}, t_{\mathtt{B}\neg\exists}, t_{\neg\mathtt{B}\exists\neg\forall}, t_{\neg\mathtt{B}\forall}\}$$

and leads straightforwardly to the obvious interpretation game. Lifting this game we derive the following fixed point interpretation strategy for the re-

ceiver under the assumption that the speaker is an expert:

$$R^* = \left\{ \begin{array}{lcl} m_{\mathtt{B}\wedge\mathtt{some}} & \mapsto & t_{[\mathtt{B}\exists\neg\forall]} \\ m_{\mathtt{B}\vee\mathtt{some}} & \mapsto & t_{[\mathtt{B}\neg\exists,\ \neg\mathtt{B}\exists\neg\forall]} \\ m_{\mathtt{B}\wedge\mathtt{all}} & \mapsto & t_{[\mathtt{B}\forall]} \\ m_{\mathtt{B}\vee\mathtt{all}} & \mapsto & t_{[\mathtt{B}\neg\exists,\ \neg\mathtt{B}\forall]} \\ m_{\mathtt{B}} & \mapsto & t_{[\mathtt{B}\neg\exists]} \\ m_{\mathtt{some}} & \mapsto & t_{[\neg\mathtt{B}\exists\neg\forall]} \\ m_{\mathtt{all}} & \mapsto & t_{[\neg\mathtt{B}\forall]} \end{array} \right\}.$$

Recall that the notation in square brackets gives information states, so that for instance, $t_{[\mathtt{B}\neg\exists,\ \neg\mathtt{B}\exists\neg\forall]}$ is an information state in which the sender thinks either of two things is possible: (i) Kai ate only broccoli or (ii) Kai ate no broccoli and some but not all of the peas. This is then the prediction of the model which also vindicates the attested implicatures in (33b)–(33e). The unattested implicature in (33f) does not follow: in state $t_{[\mathtt{B}\neg\exists,\ \neg\mathtt{B}\exists\neg\forall]}$ the sender is uncertain whether Kai had the broccoli or not.

The predictions of the model under the assumption that the sender is not an expert are rather unwieldy but not at all implausible. The fixed point interpretation strategy we derive for this case is

$$R^* = \left\{ \begin{array}{lcl} m_{\mathtt{B}\wedge\mathtt{some}} & \mapsto & t_{[\mathtt{B}\exists\neg\forall]}, t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\forall]} \\ m_{\mathtt{B}\vee\mathtt{some}} & \mapsto & t_{[\mathtt{B}\neg\exists,\neg\mathtt{B}\exists\neg\forall]},\ t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\neg\exists,\neg\mathtt{B}\exists\neg\forall]},\ t_{[\mathtt{B}\forall,\mathtt{B}\neg\exists,\neg\mathtt{B}\exists\neg\forall]}, \\ & & t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\forall,\mathtt{B}\neg\exists,\neg\mathtt{B}\exists\neg\forall]},\ t_{[\mathtt{B}\neg\exists,\neg\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\forall]}, \\ & & t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\neg\exists,\neg\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\forall]}, t_{[\mathtt{B}\forall,\mathtt{B}\neg\exists,\neg\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\forall]}, \\ & & t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\forall,\mathtt{B}\neg\exists,\neg\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\forall]} \\ m_{\mathtt{B}\wedge\mathtt{all}} & \mapsto & t_{[\mathtt{B}\forall]} \\ m_{\mathtt{B}\vee\mathtt{all}} & \mapsto & t_{[\mathtt{B}\neg\exists,\neg\mathtt{B}\forall]}, t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\neg\exists,\neg\mathtt{B}\forall]}, t_{[\mathtt{B}\forall,\mathtt{B}\neg\exists,\neg\mathtt{B}\forall]} \\ & & t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\forall,\mathtt{B}\neg\exists,\neg\mathtt{B}\forall]} \\ m_{\mathtt{B}} & \mapsto & t_{[\mathtt{B}\neg\exists]}, t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\neg\exists]}, t_{[\mathtt{B}\forall,\mathtt{B}\neg\exists]}, t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\forall,\mathtt{B}\neg\exists]} \\ m_{\mathtt{some}} & \mapsto & t_{[\neg\mathtt{B}\exists\neg\forall]},\ t_{[\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\exists\neg\forall]},\ t_{[\mathtt{B}\forall,\neg\mathtt{B}\exists\neg\forall]}, \\ & & t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\forall,\neg\mathtt{B}\exists\neg\forall]},\ t_{[\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\forall]},\ t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\forall,\neg\mathtt{B}\forall]}, \\ & & t_{[\neg\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\forall]},\ t_{[\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\forall]},\ t_{[\mathtt{B}\forall,\neg\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\forall]}, \\ & & t_{[\mathtt{B}\exists\neg\forall,\mathtt{B}\forall,\neg\mathtt{B}\exists\neg\forall,\neg\mathtt{B}\forall]} \\ m_{\mathtt{all}} & \mapsto & t_{[\neg\mathtt{B}\forall]}, t_{[\mathtt{B}\forall,\neg\mathtt{B}\forall]} \end{array} \right\}.$$

This prediction seems to exceed intuition in detail and precision, but a quick check validates that the attested implicatures in (33b)–(33e) still hold and that the unattested implicature in (33f) still does not follow from the interpretation of $m_{\mathtt{B}\vee\mathtt{some}}$. To see that (33f) does not follow, for instance, notice that each

information state associated with $m_{\mathtt{B} \vee \mathtt{some}}$ has the sender uncertain whether Kai had broccoli. Without going into the details of the derivation, suffice it to say that the IBR model does not run into the same problem as naïve scalar reasoning with a negation of (34) because it integrates the epistemic implicatures also associated with (34), namely that the speaker is uncertain between disjuncts, when comparing alternative forms.

Disjunctions Revisited

It remains to be checked whether the IBR model can also deal with disjunctions with more than two logically independent disjuncts, and with disjunctions of logically dependent disjuncts.[16]

GENERALIZED DISJUNCTIONS. Consider the case of disjunction $A \vee B \vee C$ with three logically independent disjuncts. The construction of the context model will be a mere automatism once we have settled on the correct set of alternatives to our target message $m_{\mathtt{A} \vee \mathtt{B} \vee \mathtt{C}}$. If we assume that a three-placed disjunction has no further hierarchical structure in its logical form, it is plausible to assume that at least the disjuncts $m_{\mathtt{A}}$, $m_{\mathtt{B}}$ and $m_{\mathtt{C}}$ are alternatives (cf. Sauerland 2004). With these messages our rules for canonical state construction will identify seven states:

$$T = \{t_{\mathtt{A}},\ t_{\mathtt{B}},\ t_{\mathtt{C}},\ t_{\mathtt{AB}},\ t_{\mathtt{AC}},\ t_{\mathtt{BC}},\ t_{\mathtt{ABC}}\}$$

where notation $t_{\mathtt{AB}}$, for instance, represents a state where propositions A and B are true, while C is not true. Consider a signaling game with this set of states and only the four messages above. Then, lifting the signaling game and assuming an expert sender, the forms $m_{\mathtt{A}}$, $m_{\mathtt{B}}$ and $m_{\mathtt{C}}$ will be associated with singleton information states $t_{[\mathtt{A}]}$, $t_{[\mathtt{B}]}$ and $t_{[\mathtt{C}]}$, as we would expect, but the target form $m_{\mathtt{A} \vee \mathtt{B} \vee \mathtt{C}}$ will be associated with all remaining information states because there is nothing else to express these with. That means that for proper predictions under the IBR model, the set of alternatives should include additional messages that can take away, so to speak, undesirable meanings from our target form. For the time being, let us assume that the set of alternatives for interpretation of $m_{\mathtt{A} \vee \mathtt{B} \vee \mathtt{C}}$ includes all the conjunctions of single disjuncts, as well as all disjunctions thereof (see below for discussion):

$$M = \{m_{\mathtt{A}},\ m_{\mathtt{B}},\ m_{\mathtt{C}},\ m_{\mathtt{AB}},\ m_{\mathtt{AC}},\ m_{\mathtt{BC}},\ m_{\mathtt{ABC}},\ m_{\mathtt{A,B}},\ m_{\mathtt{A,C}}, m_{\mathtt{C,B}},\ m_{\mathtt{A,B,C}}\}$$

16. Two propositions A and B are LOGICALLY INDEPENDENT iff for all $X \in \{A, \overline{A}\}$, $Y \in \{B, \overline{B}\}$ we have $X \cap Y \neq \emptyset$.

where notation $m_{\mathtt{AB}}$ stands for "A and B" and $m_{\mathtt{A,B}}$ for "A or B."

With these alternatives, the construction of the lifted and unlifted context models is nothing out of the ordinary. The model predicts that for expert senders the receiver's interpretation strategy has a unique fixed point in:

$$R^* = \left\{ \begin{array}{lcl} m_{\mathtt{A}} & \mapsto & t_{[\mathtt{A}]} \\ m_{\mathtt{B}} & \mapsto & t_{[\mathtt{B}]} \\ m_{\mathtt{C}} & \mapsto & t_{[\mathtt{C}]} \\ m_{\mathtt{AB}} & \mapsto & t_{[\mathtt{AB}]} \\ m_{\mathtt{AC}} & \mapsto & t_{[\mathtt{AC}]} \\ m_{\mathtt{BC}} & \mapsto & t_{[\mathtt{BC}]} \\ m_{\mathtt{ABC}} & \mapsto & t_{[\mathtt{ABC}]} \\ m_{\mathtt{A,B}} & \mapsto & t_{[\mathtt{A,B}]} \\ m_{\mathtt{A,C}} & \mapsto & t_{[\mathtt{A,C}]} \\ m_{\mathtt{B,C}} & \mapsto & t_{[\mathtt{B,C}]} \\ m_{\mathtt{A,B,C}} & \mapsto & t_{[\mathtt{A,B,C}]} \end{array} \right\}.$$

This is exactly as it should be. The preferred interpretation $t_{[\mathtt{A,B,C}]}$ of our target form represents an information state of the sender in which she consider it possible that either disjunct is true alone, and in which she can exclude all further possibilities.

Similarly, the predictions for the non-expert sender case are appealing, but not very perspicuous, which is why I omit obvious permutations:

$$R^* = \left\{ \begin{array}{lcl} m_{\mathtt{A}} & \mapsto & t_{[\mathtt{A}]}, t_{[\mathtt{A,AB}]}, t_{[\mathtt{A,AC}]}, t_{[\mathtt{AB,AC}]}, t_{[\mathtt{A,AB,AC}]}, t_{[\mathtt{A,ABC}]}, t_{[\mathtt{A,AB,ABC}]}, \\ & & t_{[\mathtt{A,AC,ABC}]}, t_{[\mathtt{AB,AC,ABC}]}, t_{[\mathtt{A,AB,AC,ABC}]} \\ m_{\mathtt{AB}} & \mapsto & t_{[\mathtt{AB}]}, t_{[\mathtt{AB,ABC}]} \\ m_{\mathtt{ABC}} & \mapsto & t_{[\mathtt{ABC}]} \\ m_{\mathtt{A,B}} & \mapsto & t_{[\mathtt{A,B}]}, t_{[\mathtt{A,B,AB}]}, t_{[\mathtt{A,B,AC}]}, t_{[\mathtt{A,B,AB,AC}]}, t_{[\mathtt{A,B,BC}]}, \\ & & t_{[\mathtt{A,B,AB,BC}]}, t_{[\mathtt{A,B,AC,BC}]}, t_{[\mathtt{A,B,AB,AC,BC}]}, t_{[\mathtt{A,B,ABC}]}, \\ & & t_{[\mathtt{A,B,AB,ABC}]}, t_{[\mathtt{A,B,AC,ABC}]}, t_{[\mathtt{A,B,AB,AC,ABC}]}, \\ & & t_{[\mathtt{A,B,BC,ABC}]}, t_{[\mathtt{A,B,AB,BC,ABC}]}, \\ & & t_{[\mathtt{A,B,AC,BC,ABC}]}, t_{[\mathtt{A,B,AB,AC,BC,ABC}]} \\ m_{\mathtt{A,B,C}} & \mapsto & t_{[\mathtt{A,B,C}]}, t_{[\mathtt{A,B,C,AB}]}, t_{[\mathtt{A,B,C,AC}]}, \\ & & t_{[\mathtt{A,B,C,AB,AC}]}, t_{[\mathtt{A,B,C,BC}]}, t_{[\mathtt{A,B,C,AB,BC}]}, \\ & & t_{[\mathtt{A,B,C,AC,BC}]}, t_{[\mathtt{A,B,C,AB,AC,BC}]}, t_{[\mathtt{A,B,C,ABC}]}, \\ & & t_{[\mathtt{A,B,C,AB,ABC}]}, t_{[\mathtt{A,B,C,AC,ABC}]}, t_{[\mathtt{A,B,C,AB,AC,ABC}]}, \\ & & t_{[\mathtt{A,B,C,BC,ABC}]}, t_{[\mathtt{A,B,C,AB,BC,ABC}]}, t_{[\mathtt{A,B,C,AC,BC,ABC}]}, \\ & & t_{[\mathtt{A,B,C,AB,AC,BC,ABC}]} \end{array} \right\}$$

Our target form is interpreted as conveying that the sender is in an epistemic state in which she considers it possible that A, B, and C are all true alone, and that she might consider further alternatives possible, but that she does not have any more concrete knowledge beyond that.

The predictions based on the above set M are flawless, but the question remains whether the set M itself is defensible. Conceptually, M is just what we would obtain from an application of, for instance, Sauerland's (2004) construction of alternatives to disjunction, where we neglect any internal binary structure of disjunction and treat it as flat, at least on the level of logical form.

Moreover, we do not necessarily need exactly this set of alternatives, if we are only interested in the interpretation of the target form. More concretely, the situation for the IBR model is this. We could in principle assume a smaller set of alternatives, namely

$$M' = \{m_{\mathtt{A}},\ m_{\mathtt{B}},\ m_{\mathtt{C}},\ m_{\mathtt{A,B}},\ m_{\mathtt{A,C}}, m_{\mathtt{C,B}},\ m_{\mathtt{A,B,C}}\}$$

and still the predictions for the target message $m_{\mathtt{A,B,C}}$ would be the exact same: only for the non-expert case would the predictions for non-target disjunctive messages $m_{\mathtt{A,B}}$ change to include more possible interpretations that could otherwise be expressed by more specific conjunctions. Similarly, we could also take a much larger set, namely the smallest set including $m_{\mathtt{A}}$, $m_{\mathtt{B}}$ and $m_{\mathtt{C}}$ which is closed under disjunction and conjunction. Again, the prediction for the target message alone will be flawless, but non-target messages, in particular, disjunctions with entailing disjuncts of the form "A or (A and B)" are not assigned intuitive interpretations. This is a separate problem that I will be dealing with presently. As far as the interpretation of disjunctions with more than two mutually exclusive disjuncts is concerned, we should conclude that the IBR model can deal with those under several possible specifications of alternative sets.

The principles of model specification discussed here for three-place disjunctions with mutually exclusive disjuncts should generalize to disjunctions with an arbitrary number of disjuncts. The predictions discussed here are based on simulation, but it would be desirable to offer an analytic result stating the predictions of the IBR model for any arbitrary disjunction. Unfortunately, this interesting issue has to be left for future research.

Entailing Disjuncts. A problem of interpretation may arise under truth-conditional semantics for disjunctions of the form "A or (A and B)", or more generally whenever one disjunct entails another. The problem is that, as far as

truth-conditions carry us, 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 an expert on the topic at hand. Compare the answers (37a) and (38a) to a question (36).

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

(37) a. John did.

b. $\rightsquigarrow$ The speaker knows that John came and that Mary did not.

(38) a. John or (John and Mary).

b. $\rightsquigarrow$ The speaker knows that John came and considers it possible that Mary came too.

Though semantically equivalent, the implicatures associated with these answers, (37b) and (38b) respectively, are clearly different. This problem palpably affects pretty much all global Neo-Gricean accounts that rely on truth-conditional semantics, and it is thus interesting to see what the present GTP perspective can add to this puzzle.

Let us therefore derive a canonical context model from the target expression "A or (A and B)." Following the construction of alternatives for disjunction from before, we take all individual disjuncts occurring in the target expression, $m_{\mathtt{A}}$ and $m_{\mathtt{AB}}$, and also all conjunctions and disjunctions thereof. That way we obtain a set of alternatives:

$$M = \{m_{\mathtt{A}},\ m_{\mathtt{B}},\ m_{\mathtt{A}\vee\mathtt{AB}},\ m_{\mathtt{A}\wedge\mathtt{AB}}\}\,.$$

These alternatives yield only a binary state distinction, differentiating a state $t_{\mathtt{A}}$, where only A is true, from a state $t_{\mathtt{AB}}$, where A and B are both true:

$$T = \{t_{\mathtt{A}}, t_{\mathtt{AB}}\}\,.$$

Obviously, on this simple set of states not only is $m_{\mathtt{A}\vee\mathtt{AB}}$ equivalent to $m_{\mathtt{A}}$, but also $m_{\mathtt{A}\wedge\mathtt{AB}}$ is equivalent to $m_{\mathtt{B}}$. If we stick to truth-conditional meaning only, it is impossible to distinguish between either of these linguistic forms in the model based on their meaning. If we nonetheless want to include a non-semantic distinction, the most obvious idea is to assume that the syntactically more complex forms are more costly than their semantically equivalent correspondences: I suggest that nominal message costs should be brought in whenever there is an obvious morpho-syntactic difference between two semantically equivalent forms (compare also M-implicatures).

Feeding this assumption into the canonical interpretation game, and lifting the model, the expert case yields the fixed point interpretation:

$$R^* = \left\{ \begin{array}{lcl} m_{\mathtt{A}} & \mapsto & t_{[\mathtt{A}]} \\ m_{\mathtt{AB}} & \mapsto & t_{[\mathtt{AB}]} \\ m_{\mathtt{A} \vee \mathtt{AB}} & \mapsto & t_{[\mathtt{A},\mathtt{AB}]} \\ m_{\mathtt{A} \wedge \mathtt{AB}} & \mapsto & \text{surprise} \end{array} \right\}.$$

This is exactly how we would want it to be. Only the interpretation of the non-expert case calls for careful thought:

$$R^* = \left\{ \begin{array}{lcl} m_{\mathtt{A}} & \mapsto & t_{[\mathtt{A}]}, t_{[\mathtt{A},\mathtt{AB}]} \\ m_{\mathtt{AB}} & \mapsto & t_{[\mathtt{AB}]} \\ m_{\mathtt{A} \vee \mathtt{AB}} & \mapsto & \text{surprise} \\ m_{\mathtt{A} \wedge \mathtt{AB}} & \mapsto & \text{surprise} \end{array} \right\}.$$

The prediction here is that the target message comes as a surprise. This is not at all unreasonable, because associating $t_{[\mathtt{A},\mathtt{AB}]}$ with $m_{\mathtt{A}}$ is crucially what an interpretation of $m_{\mathtt{A}}$ under an inexpert assumption entails. Extrapolating from this, we predict that the costly form $m_{\mathtt{A} \vee \mathtt{AB}}$ could be optimal for an inexpert sender, for instance, if she believed that the receiver incorrectly assumed her to be an expert. (This kind of speaker uncertainty about the receiver's assumption about speaker expertise is not modelled explicitly, but it is easy to see how it could be.)

Summary. Taking stock, in this section I have suggested that the basic signaling game model, in which the sender knows the true state of the world, can accommodate the speaker's partial information by a conceptual reinterpretation of the notion of a state. In order to derive epistemic implicatures, as I have called them, we should:

(i) lift the interpretation of states in the game model from states of the world to information states of the sender;

(ii) lift the notion of semantic meaning captured in $[\![\cdot]\!]$ from "being true in a state of the world" to "believed true in an information state";

(iii) implement a speaker competence assumption in $\Pr(\cdot)$ such that smaller, i.e., more specific, information states are more probable if the speaker is assumed to be an expert;

(iv) stick to an interpretation game with $T = A$ and corresponding payoffs.

I showed how this basic set-up deals with the epistemic implicatures of disjunctions and also with the McCawley-Chierchia problem. In the following, I will show how the IBR model, if applied to lifted signaling games, also copes with free choice inferences.

3.3 Free Choice Inferences

A linguistically very interesting case concerning the meaning and use of English disjunction is its interaction with deontic modals such as in (39a) and (40a) for which we obtain the FREE-CHOICE READINGS (FC-reading) in (39b) and (40b).

(39) a. You may take an apple or a pear. $\Diamond(A \vee B)$

b. *Free Choice Implicature:* You may take an apple and you may take a pear. $\Diamond(A) \wedge \Diamond(B)$

c. *Scalar Implicature:* You may not take both. $\neg\Diamond(A \wedge B)$

(40) a. You must take an apple or a pear. $\Box(A \vee B)$

b. *Free Choice Implicature:* You may take an apple and you may take a pear. $\Diamond(A) \wedge \Diamond(B)$

c. *Scalar Implicature:* You need not take both. $\neg\Box(A \wedge B)$

The basic observation (see Kamp 1973, 1978) is that, contrary to what we might expect from a standard logical semantics of modals and disjunction, sentences like (39a) and (40a) give rise to the *conjunctive* reading in (39b) and (40b) under the scope of the deontic modals, and the strengthened *exclusive* readings in (39c) and (40c). In the following, I am especially concerned with the former FC-readings, but I will also deal with the latter exclusive readings on the side. I will adopt the mainstream conviction in the linguistic literature that both aspects of meaning are not part of the standard semantic meaning of sentences (39a) and (40a), and should rather be accounted for as conversational implicatures. Let us briefly revise the relevant arguments behind the mainstream conviction.

A standard possible-worlds interpretation of deontic modals renders $\Diamond A$ ($\Box A$) true in a pointed model $\langle W, R, w\rangle$, with a deontic accessibility relation $R \subseteq W \times W$, if some (all) worlds accessible from w via R make the proposition A true. Under these semantics, $\Diamond(A \vee B)$ is true in a pointed model $\langle W, R, w\rangle$ iff some worlds R-accessible from w make the disjunction $A \vee B$

true. In particular then, the standard semantics renders (39a) and (40a) true in pointed models in which the intuitively attested inferences (39b) and (39c), respectively (40b) and (40c), do *not* hold. To see this, both $\Diamond(A \vee B)$ and $\Box(A \vee B)$ are true in a pointed model in which A and only A is true for all accessible worlds. In these pointed models however the FC-reading (39b) and (40b) that the hearer may choose between alternatives A and B is not true. Similarly, both $\Diamond(A \vee B)$ and $\Box(A \vee B)$ are true in a pointed model in which both A and B are true for all accessible worlds. In these pointed models, the exclusive readings in (39c) and (40c) are not true. Taken together, both the FC-reading and the exclusive readings would require a strengthening of the standard semantics of modals and disjunction.

Reassessing Disjunction. Of course, one may argue that the standard semantics is wrong and that it needs to be strengthened accordingly. Alternative semantics for disjunction that do exactly that have been suggested by Zimmermann (2000), Geurts (2005) and Simons (2005). Against an amendment of the semantics of disjunction, other authors have argued that the intuitive readings we are after should rather be accounted for as conversational implicatures (Kratzer and Shimoyama 2002; Alonso-Ovalle 2005; Schulz 2005). One argument voiced in favor of this position is that the attested FC-reading does not (necessarily) arise in downward entailing contexts such as (41).

(41) No one is allowed to take an apple or a pear.

Another argument in favor of an implicature-based analysis of FC-readings is the observation that the FC-inferences in (39b) and (40b) seem 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 the following example (42a) we do not get the FC-implicature. We rather get the *ignorance implicature* (42b) similar to the one we got for a plain disjunction. Additionally, the epistemic ignorance reading of $\Diamond(A \vee B)$ forced in (42a) may still convey an (epistemic) scalar implicature as in (42c).

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

b. *Ignorance Implicature:* The speaker doesn't know whether the hearer may take an apple and the speaker does not know whether the hearer may take a pear. $\neg K_S(\Diamond A) \wedge \neg K_S(\Diamond B)$

c. *Scalar Implicature:* The speaker knows that the receiver may not take both. $K_S \neg \Diamond(A \wedge B)$

Reassessing Deontic Modals. A similarly non-classical approach to FC-readings might reconsider the analysis of deontic modals. In this camp, we find the earliest contribution on the subject (Kamp 1973), and thereafter Merin (1992) and van Rooij (2000) who all have favored a *performative analysis* of the deontic modals in FC-environments. A general conceptual problem for a performative analysis is that FC-readings also arise for modals which are clearly *not* performatively used. Whence that all else being equal a uniform analysis should be preferred (see Schulz 2005, for this line of argument).[17]

Taken together, the intuitively attested FC-readings pose a problem for standard semantics of disjunction and deontic modals. Deviations from the classical, logical analyses are on the market, but obviously the preferred Gricean analysis would adhere to the standards and give a rationalistic account of the readings in question.

3.3.1 Free Choice from Anti-Exhaustivity

Unfortunately, FC-inferences cannot be derived by naïve scalar reasoning unless we buy into additional and perhaps seemingly *ad hoc* assumptions about the nature of the relevant alternative forms to reason with. To see this, consider the set M_1 which I and others take to be the most natural set of alternatives to a disjunctive permission $\Diamond(A \vee B)$ (see Kratzer and Shimoyama 2002; Chierchia 2004; Alonso-Ovalle 2005; Fox 2007; Chierchia et al. 2008):[18]

$$M_1 = \left\{ m_{\Diamond \mathtt{A}}, m_{\Diamond \mathtt{B}}, m_{\Diamond(\mathtt{A} \vee \mathtt{B})}, m_{\Diamond(\mathtt{A} \wedge \mathtt{B})} \right\},$$

which consists of the sentences in (43).

(43) a. You may take an apple. $m_{\Diamond \mathtt{A}}$

b. You may take a pear. $m_{\Diamond \mathtt{B}}$

c. You may take an apple or a pear. $m_{\Diamond(\mathtt{A} \vee \mathtt{B})}$

d. You may take an apple and a pear. $m_{\Diamond(\mathtt{A} \wedge \mathtt{B})}$

17. Yet a different kind of analysis is pursued by Asher and Bonevac (2005), who argue that a permission statement "it's permitted that A" should be analyzed as a defeasible conditional "if you do A (and everything is normal), it is okay." I will come back briefly to this proposal in section 3.3.3.

18. Notice that I am presently only concerned with FC-readings and not the inferences from "may" to "not have to." That is why, like many others before me, I only look at alternatives to disjunction. (I would like to thank Maria Aloni for raising this issue.)

It is not difficult to see how things go wrong for naïve scalar reasoning based on the set M_1. If we assume that an assertion of a form X implicates that all stronger alternatives for X from the relevant set of candidate forms are false, we derive that an utterance of $\Diamond(A \vee B)$ implicates $\neg\Diamond A$ and $\neg\Diamond B$, which is clearly too strong, and even incompatible in conjunction with the semantic meaning of the asserted sentence.[19] So, given these allegedly natural alternatives a naïve account cannot derive the FC-reading.

OTHER ALTERNATIVES. The mistake may be sought in the set of featured alternatives, of course. And indeed, other authors have featured other alternatives. For instance, Schulz (2005) and Aloni and van Rooij (2007) use the following set of alternatives:

$$M_2 = \left\{ m_{\Box \mathtt{A}}, m_{\Box \mathtt{B}}, m_{\Box(\neg \mathtt{A})}, m_{\Box(\neg \mathtt{B})} \right\}$$

and derive the desired FC-inference. It's clear however that this choice of alternatives stands in need of justification; at least more so than did the previous. All else being equal, I believe, we would prefer an account that derives FC-readings as implicatures from the set M_1 rather than from the set M_2.

ANTI-EXHAUSTIVITY. This seems indeed possible, as an interesting observation by Kratzer and Shimoyama (2002) suggests: Kratzer and Shimoyama hypothesize that one reason why $m_{\Diamond(\mathtt{A} \vee \mathtt{B})}$ may be used instead of $m_{\Diamond \mathtt{A}}$ and $m_{\Diamond \mathtt{B}}$ is so as *to prevent an exhaustive interpretation* of the latter forms.[20] Since, in other words, on this account the purpose of using $m_{\Diamond(\mathtt{A} \vee \mathtt{B})}$ is to negate an exhaustive inference, the term ANTI-EXHAUSTIVITY has caught hold in the community to describe Kratzer and Shimoyama's idea.

The desired prediction of free choice follows from naïve scalar reasoning if we assume that the alternative forms $m_{\Diamond \mathtt{A}}$ and $m_{\Diamond \mathtt{B}}$ in M_1 should be interpreted *exhaustively*, i.e., basically as in (44a) and (44b) respectively.

19. For concreteness' sake, let me mention that an epistemic version of this problem arises for instance for Sauerland (2004)'s improved Neo-Gricean model of implicature calculation. As Fox (2007) notes critically, Sauerland (2004) predicts that an utterance of $\Diamond(A \vee B)$ implicates that the speaker does not know that $\Diamond A / \Diamond B$. This is indeed too strong a prediction for the attested FC-reading (although it is the correct prediction for the epistemic reading in (42a)).

20. More concretely, Kratzer and Shimoyama (2002) offer a pragmatic account of the use of German existential 'irgendein' which they analyze as introducing domain widening. Domain widening under an existential modal can then be rationalized parallel to the use of disjunction.

(44) a. You may take an apple but you may not take a pear. $m_{\Diamond \mathtt{A} \wedge \neg \Diamond \mathtt{B}}$

b. You may take a pear but you may not take an apple. $m_{\Diamond \mathtt{B} \wedge \neg \Diamond \mathtt{A}}$

Effectively, this amounts to replacing the set of alternatives M_1 with the set:

$$M_1^* = \left\{ m_{\Diamond \mathtt{A} \wedge \neg \Diamond \mathtt{B}}, m_{\Diamond \mathtt{B} \wedge \neg \Diamond \mathtt{A}}, m_{\Diamond (\mathtt{A} \vee \mathtt{B})}, m_{\Diamond (\mathtt{A} \wedge \mathtt{B})} \right\}.$$

If we now apply the standard Neo-Gricean mechanism of implicature calculation, we derive that an utterance of $\Diamond(A \vee B)$ implicates that all stronger alternatives in M_1^* are false, i.e., we derive the implicatures in (45).

(45) a. It's not the case that the hearer may take an apple but not a pear.

b. It's not the case that the hearer may take a pear but not an apple.

Truth of $\Diamond(A \vee B)$ in conjunction with the implicatures in (45) derives the FC-reading, as the interested reader will quickly be able to verify.

HIDDEN EXHAUSTIVE OPERATORS. So far, so good. The desired FC-readings follow from anti-exhaustivity, but where does anti-exhaustivity itself find its legitimation? Chierchia (2004), Fox (2007) and Chierchia et al. (2008) answer this question by appeal to their general theory of *local* implicature calculation *in the syntax*: the gist of the idea is that a hidden exhaustivity operator —akin to the meaning of "only"— applies in the syntax, if necessary multiple times, to supply the proper readings of alternatives and target forms. Without going into the details of any individual account, suffice it to note that this grammatical approach sticks with the original set of alternatives M_1 from which it derives M_1^* by insertion of hidden exhaustivity operators at the required places in the syntactic derivation of (39a). Further applications of exhaustivity operators —higher up in the syntactic derivation— would then feed on the alternative set M_1^* and derive the FC-reading along the lines spelled out above.

The reader will find this latter localist account of FC-inferences appealing to the extent that she is open towards the somewhat iconoclastic idea of relegating basic pragmatic mechanisms to syntax; conversely, she will dislike the suggested solution proportional to her sense that reiterations of hidden syntactic operators not only create heavy theoretical overload, but also seem rather unwieldy and arbitrary, at least compared to the strong folk-psychologicl appeal of the Gricean rationalistic programme. But the ball is in the field of the classical Griceans to meet the challenge posed by the syntax-enthusiasts who write:

> "We believe this logic [i.e., the logic of anti-exhaustivity] is basically correct, but we don't see a way to derive it from basic principles of communication (Maxims). [...] In conclusion, we have sketched reasons to believe that free choice effects can be explained in a principled way as meta- (or higher order) implicatures. If this is anywhere close to the mark, then clearly implicatures must be part of grammar."
>
> (Chierchia et al. 2008, p. 36)

In the following I would like to rise to this challenge and show how the IBR model can account for the FC-readings of (39a) based on the alternatives in M_1 pretty much by deriving anti-exhaustivity from iteration: peeking ahead, it will turn out that early iteration steps derive exhaustive readings of forms $m_{\Diamond \mathtt{A}}$ and $m_{\Diamond \mathtt{B}}$; later iterations will then compare $m_{\Diamond(\mathtt{A} \vee \mathtt{B})}$ with the exhaustive interpretations of $m_{\Diamond \mathtt{A}}$ and $m_{\Diamond \mathtt{B}}$. Iteration thus implements the logic of anti-exhaustivity, and explains FC-readings in rationalistic terms without relegating implicatures to syntax.

3.3.2 Anti-Exhaustivity from Iteration

In this section I would like to spell out how FC-readings and also ignorance readings can be derived in the IBR model. Vital for my account is a proper defense of the assumptions feeding the construction of a reasonable signaling game model. The present account does not require any special assumptions beyond the general principles for the construction of interpretation games that I defended in section 3.1.1. The only thing that deserves motivation is my use of non-lifted and lifted models to account for FC-readings and ignorance readings respectively. This is what I will do first.

Authorities and Experts

A sentence like (39a) basically allows for two kinds of readings: the FC-reading in (39b), and the ignorance reading in (42b). (This is similar for the universal deontic modal in (40a), of course.) There is a strong intuition that the reading we obtain depends on how well-informed we take the speaker to be: where it is the speaker herself who is the relevant authority responsible for granting or withholding permission, FC-readings arise; where the speaker appears at best a possibly underinformed reporter on the deontic state of affairs, ignorance readings arise.

To make this intuition bite in the game theoretic context model I will distinguish terminologically *(deontic) authorities* from *(epistemic) experts*. Authorities

are (assumed) infallible informants who can*not* err when describing the relevant deontic states of affairs.[21] Experts, on the other hand, may also happen to be perfectly informed, but they are not the ultimate authority so that error is at least in principle conceivable. In a nutshell: experts on deontic matters may be mistaken, authorities cannot.

Consequently, I propose to model the context of utterance either as a non-lifted or as a lifted signaling game. If the speaker is (assumed to be) an authority, the context model will be a normal, i.e., *non-lifted* signaling game where it is common belief that the sender knows the true state of the world. The states of the unlifted game model fix which of the actions A (taking an apple) and B (taking a pear) are feasible or allowed actions for the hearer. In contrast, if the speaker is not an absolute authority, the context model will be an *epistemically lifted* signaling game in which the sender may (or may not) have imperfect information about the deontic state of affairs. The states in the lifted game are thus information states representing the speaker's information concerning what obligations and permissions obtain. Unlike in the non-lifted game, the sender is not assumed to necessarily be perfectly informed.

Both context models, basic and lifted, are games with interpretation actions, so as to clearly model the pragmatic inferences about the meaning of the sentences involved.[22] Based on this, the next section will show how the basic, non-lifted models should derive FC-readings for both "may" and "must." The subsequent section covers the ignorance readings for both modals.

Deriving FC-Readings

Existential Modals. As for the case "may(A or B)," I suggest to adopt the following non-lifted model from which the lifted model will be derived later on. We would like to stick to the arguably most natural set of alternatives, as discussed in section 3.3.1:

$$M = \left\{ m_{\Diamond \mathtt{A}}, m_{\Diamond \mathtt{B}}, m_{\Diamond (\mathtt{A} \vee \mathtt{B})}, m_{\Diamond (\mathtt{A} \wedge \mathtt{B})} \right\}.$$

21. Notice that I still adhere to a descriptive approach: authorities still describe the deontic states of affairs; they do not performatively create, remove or change obligations by sending messages.

22. If we assume that an interpretation game is played on this set of states, we are basically construing the context of utterance for (39a) as one in which the (implicit) question under discussion is: "which combination of actions out of $\{A, B\}$ may I (the receiver) perform so as to please you (the sender)?" Alternatively, we could have the receiver respond by performing the concrete actions A and B. I stick to the interpretation framework for continuity with previous and subsequent cases.

	$\Pr(t)$	$t_{\Diamond \mathtt{A}}$	$t_{\Diamond \mathtt{B}}$	$t_{\Diamond \mathtt{AB}}$	$t_{\Diamond \mathtt{A}\mid\mathtt{B}}$	$m_{\Diamond \mathtt{A}}$	$m_{\Diamond \mathtt{B}}$	$m_{\Diamond(\mathtt{A}\vee\mathtt{B})}$	$m_{\Diamond(\mathtt{A}\wedge\mathtt{B})}$
$t_{\Diamond \mathtt{A}}$	$1/4$	1,1	0,0	0,0	0,0	$\surd$	–	$\surd$	–
$t_{\Diamond \mathtt{B}}$	$1/4$	0,0	1,1	0,0	0,0	–	$\surd$	$\surd$	–
$t_{\Diamond \mathtt{AB}}$	$1/4$	0,0	0,0	1,1	0,0	$\surd$	$\surd$	$\surd$	$\surd$
$t_{\Diamond \mathtt{A}\mid\mathtt{B}}$	$1/4$	0,0	0,0	0,0	1,1	$\surd$	$\surd$	$\surd$	–

Figure 3.6: Unlifted signaling game for free choice "may(A or B)"

By our general construction rule, we then derive four states of the signaling game model under a standard possible worlds semantics:

	$m_{\Diamond \mathtt{A}}$	$m_{\Diamond \mathtt{B}}$	$m_{\Diamond(\mathtt{A}\wedge\mathtt{B})}$
$t_{\Diamond \mathtt{AB}}$	$\surd$	$\surd$	$\surd$
$t_{\Diamond \mathtt{A}\mid\mathtt{B}}$	$\surd$	$\surd$	–
incons.	$\surd$	–	$\surd$
$t_{\Diamond \mathtt{A}}$	$\surd$	–	–
incons.	–	$\surd$	$\surd$
$t_{\Diamond \mathtt{B}}$	–	$\surd$	–
incons.	–	–	$\surd$
incons.	–	–	–

The state $t_{\Diamond \mathtt{AB}}$, for instance, is one where the receiver is allowed to take both an apple and a pear. The state $t_{\Diamond \mathtt{A}\mid\mathtt{B}}$, on the other hand, is one where the receiver may take either an apple or a pear but not both. With our usual assumptions of flat priors and cheap talk, we thus arrive at the signaling game in figure 3.6.

For this game, the IBR model predicts a unique fixed point interpretation behavior of the receiver for both strands of reasoning:

$$R^* = \left\{ \begin{array}{lcl} m_{\Diamond \mathtt{A}} & \mapsto & t_{\Diamond \mathtt{A}} \\ m_{\Diamond \mathtt{B}} & \mapsto & t_{\Diamond \mathtt{B}} \\ m_{\Diamond(\mathtt{A}\vee\mathtt{B})} & \mapsto & t_{\Diamond \mathtt{A}\mid\mathtt{B}} \\ m_{\Diamond(\mathtt{A}\wedge\mathtt{B})} & \mapsto & t_{\Diamond \mathtt{AB}} \end{array} \right\}.$$

This is the desired prediction for interpretation of $m_{\Diamond(\mathtt{A}\vee\mathtt{B})}$. In order to show how in particular iteration of best response reasoning accounts for anti-exhaustivity, let me spell out and comment on the R_0-sequence for illustration.

Under the assumed semantics the naïve receiver behavior is:

$$R_0 = \left\{ \begin{array}{lcl} m_{\Diamond \mathtt{A}} & \mapsto & t_{\Diamond \mathtt{A}}, t_{\Diamond \mathtt{AB}}, t_{\Diamond \mathtt{A}|\mathtt{B}} \\ m_{\Diamond \mathtt{B}} & \mapsto & t_{\Diamond \mathtt{B}}, t_{\Diamond \mathtt{AB}}, t_{\Diamond \mathtt{A}|\mathtt{B}} \\ m_{\Diamond (\mathtt{A} \vee \mathtt{B})} & \mapsto & T \\ m_{\Diamond (\mathtt{A} \wedge \mathtt{B})} & \mapsto & t_{\Diamond \mathtt{AB}} \end{array} \right\}.$$

Based on this, the optimal strategy for the sender is:

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

It is noteworthy here that $m_{\Diamond \mathtt{A}}$ and $m_{\Diamond \mathtt{B}}$ are the best sender choices in $t_{\Diamond \mathtt{A}|\mathtt{B}}$, because under R_0's interpretation these messages yield a chance of $^1/_3$ of successful communication, as opposed to a chance of $^1/_4$ when sending $m_{\Diamond (\mathtt{A} \vee \mathtt{B})}$. Our target form will therefore be a surprise message to R_2:

$$R_2 = \left\{ \begin{array}{lcl} m_{\Diamond \mathtt{A}} & \mapsto & t_{\Diamond \mathtt{A}} \\ m_{\Diamond \mathtt{B}} & \mapsto & t_{\Diamond \mathtt{B}} \\ m_{\Diamond (\mathtt{A} \vee \mathtt{B})} & \mapsto & \text{surprise} \\ m_{\Diamond (\mathtt{A} \wedge \mathtt{B})} & \mapsto & t_{\Diamond \mathtt{AB}} \end{array} \right\}$$

μ_2	$t_{\Diamond \mathtt{A}}$	$t_{\Diamond \mathtt{B}}$	$t_{\Diamond \mathtt{AB}}$	$t_{\Diamond \mathtt{A}\|\mathtt{B}}$
$m_{\Diamond \mathtt{A}}$	$^2/_3$	0	0	$^1/_3$
$m_{\Diamond \mathtt{B}}$	0	$^2/_3$	0	$^1/_3$
$m_{\Diamond (\mathtt{A} \vee \mathtt{B})}$	0	0	0	0
$m_{\Diamond (\mathtt{A} \wedge \mathtt{B})}$	0	0	1	0

Under the vanilla model, without forward induction assumption, R_2 would respond to $m_{\Diamond (\mathtt{A} \vee \mathtt{B})}$ with any action in T. This interpretation will settle on the desired outcome eventually, as the interested reader will happily verify. Still, we can also use a shortcut, for the sake of exposition, and notice that our target message $m_{\Diamond (\mathtt{A} \vee \mathtt{B})}$ is actually weakly 2-dominated in all states except $t_{\Diamond \mathtt{A}|\mathtt{B}}$: intuitively speaking, all other states already have a message which expresses these states at that point. So, by forward induction reasoning, R_2 may arrive at the interpretation $R_2(m_{\Diamond (\mathtt{A} \vee \mathtt{B})}) = \{t_{\Diamond \mathtt{A}|\mathtt{B}}\}$, which yields the fixed point of this reasoning sequence.

Let me stress again for clarity that the predictions of the model do not hinge on forward induction. The reasoning with weak k-dominance is merely more compact, and eases exposition and helps focus on localizing the formal counterpart of the "anti-exhaustivity reasoning" in the model. Anti-exhaustivity occurs, so to speak, because R_2 interprets the forms $m_{\Diamond \mathtt{A}}$ and $m_{\Diamond \mathtt{B}}$ exhaustively as denoting states $t_{\Diamond \mathtt{A}}$ and $t_{\Diamond \mathtt{B}}$ respectively, and because R_2

compares this exhaustive interpretation to the target expression. This is so even though both forms also get sent in $t_{\Diamond \mathtt{A}|\mathtt{B}}$ by S_1, since by proper sophisticated updating, the posterior probability of $t_{\Diamond \mathtt{A}}$ after observing $m_{\Diamond \mathtt{A}}$ is twice as high as that of $t_{\Diamond \mathtt{A}|\mathtt{B}}$. In effect, the IBR model derives the intuitively appealing logic of anti-exhaustivity by a two-step iteration process: first we derive the exhaustive interpretation of $m_{\Diamond \mathtt{A}}$ and $m_{\Diamond \mathtt{B}}$, and from that arrive at the attested FC-reading. (Similar remarks apply to the S_0-sequence.)

UNIVERSAL MODALS. The present account of FC-readings carries over to universal modals without any further complications. If we assume the set of speaker alternatives:

$$M = \left\{ m_{\Box \mathtt{A}},\ m_{\Box \mathtt{B}},\ m_{\Box(\mathtt{A} \vee \mathtt{B})},\ m_{\Box(\mathtt{A} \wedge \mathtt{B})} \right\}$$

we derive four possible state distinctions:

	$m_{\Box \mathtt{A}}$	$m_{\Box \mathtt{B}}$	$m_{\Box(\mathtt{A} \wedge \mathtt{B})}$
$t_{\Box \mathtt{AB}}$	$\surd$	$\surd$	$\surd$
incons.	$\surd$	$\surd$	$-$
incons.	$\surd$	$-$	$\surd$
$t_{\Box \mathtt{A}}$	$\surd$	$-$	$-$
incons.	$-$	$\surd$	$\surd$
$t_{\Box \mathtt{B}}$	$-$	$\surd$	$-$
incons.	$-$	$-$	$\surd$
$t_{\Box \mathtt{A}|\mathtt{B}}$	$-$	$-$	$-$

Here the state $t_{\Box \mathtt{AB}}$, for instance, is one where the receiver has to take both an apple and a pear. The state $t_{\Box \mathtt{A}|\mathtt{B}}$, on the other hand, is one where the receiver has to take an apple or a pear but may choose which one. This yields the unlifted cheap talk signaling game in figure 3.7.

For this game, the IBR model again predicts a single unique receiver interpretation strategy for both strands of reasoning:

$$R^* = \left\{ \begin{array}{lcl} m_{\Box \mathtt{A}} & \mapsto & t_{\Box \mathtt{A}} \\ m_{\Box \mathtt{B}} & \mapsto & t_{\Box \mathtt{B}} \\ m_{\Box(\mathtt{A} \vee \mathtt{B})} & \mapsto & t_{\Box \mathtt{A}|\mathtt{B}} \\ m_{\Box(\mathtt{A} \wedge \mathtt{B})} & \mapsto & t_{\Box \mathtt{AB}} \end{array} \right\}.$$

The target expression receives the interpretation that the receiver need not take an apple, and that he need not take a pear, just as intuition demands.

	$\Pr(t)$	$t_{\Box \mathtt{A}}$	$t_{\Box \mathtt{B}}$	$t_{\Box \mathtt{AB}}$	$t_{\Box \mathtt{A}\mid\mathtt{B}}$	$m_{\Box \mathtt{A}}$	$m_{\Box \mathtt{B}}$	$m_{\Box(\mathtt{A}\vee\mathtt{B})}$	$m_{\Box(\mathtt{A}\wedge\mathtt{B})}$
$t_{\Box \mathtt{A}}$	$1/4$	1,1	0,0	0,0	0,0	$\surd$	–	$\surd$	–
$t_{\Box \mathtt{B}}$	$1/4$	0,0	1,1	0,0	0,0	–	$\surd$	$\surd$	–
$t_{\Box \mathtt{AB}}$	$1/4$	0,0	0,0	1,1	0,0	$\surd$	$\surd$	$\surd$	$\surd$
$t_{\Box \mathtt{A}\mid\mathtt{B}}$	$1/4$	0,0	0,0	0,0	1,1	–	–	$\surd$	–

Figure 3.7: Unlifted signaling game for free choice "must(A or B)"

Reflection. Taken together, an account of FC-readings is rather straightforward in the IBR model. Once we have settled on an acceptable set of alternative forms, we do not have to assume message costs or any particular ordering on states to derive the FC-readings. This is what sets the present approach apart from previous accounts, such as in terms of bidirectional optimality theory (Aloni 2007; Pauw 2008), or "minimal models" (Schulz 2005) or "default interpretations" (Asher and Bonevac 2005). The key to the success of IBR is, in a manner of speaking, the proper exploitation of semantic structure by sophisticated updating (as given in an assumed set of alternatives): previous accounts have not drawn on the full 'proportional information' given, so to speak, when comparing the meaning of expressions; therefore previous accounts had to rely on additional ordering assumptions.

Deriving Ignorance Implicatures

Thus far we have derived the FC-implicatures of sentences (39a) and (40a). In order to do so, we have set up a context model that modeled the deontic authority of the speaker in terms of an unlifted signaling game in which the sender cannot possibly be mistaken about the actually obtaining deontic state of affairs. Turning to the epistemic inferences associated with a situation where the speaker may in principle fail to be the absolute authority, we should try lifting the basic models. Epistemic lifting of these context models is a plain execution of the principles set out in section 3.2.

Existential Modals. Lifting the basic signaling game in figure 3.6, we receive a total of fifteen epistemic states:

$$\begin{aligned} T \;=\; \{&t_{[\Diamond \mathtt{A}]},\ t_{[\Diamond \mathtt{B}]},\ t_{[\Diamond \mathtt{A},\Diamond \mathtt{B}]},\ t_{[\Diamond \mathtt{AB}]},\ t_{[\Diamond \mathtt{A},\Diamond \mathtt{AB}]},\ t_{[\Diamond \mathtt{B},\Diamond \mathtt{AB}]},\ t_{[\Diamond \mathtt{A},\Diamond \mathtt{B},\Diamond \mathtt{AB}]},\ t_{[\Diamond \mathtt{A}\mid\mathtt{B}]},\\ &t_{[\Diamond \mathtt{A},\Diamond \mathtt{A}\mid\mathtt{B}]},\ t_{[\Diamond \mathtt{B},\Diamond \mathtt{A}\mid\mathtt{B}]},\ t_{[\Diamond \mathtt{A},\Diamond \mathtt{B},\Diamond \mathtt{A}\mid\mathtt{B}]},\ t_{[\Diamond \mathtt{AB},\Diamond \mathtt{A}\mid\mathtt{B}]},\ t_{[\Diamond \mathtt{A},\Diamond \mathtt{AB},\Diamond \mathtt{A}\mid\mathtt{B}]},\\ &t_{[\Diamond \mathtt{B},\Diamond \mathtt{AB},\Diamond \mathtt{A}\mid\mathtt{B}]},\ t_{[\Diamond \mathtt{A},\Diamond \mathtt{B},\Diamond \mathtt{AB},\Diamond \mathtt{A}\mid\mathtt{B}]}\}. \end{aligned}$$

The notation here is as before: commas in brackets separate unlifted states as epistemic possibilities not ruled out in an epistemic state. For example, the state $t_{[\Diamond \mathtt{A}, \Diamond \mathtt{B}]}$ contains the unlifted states $t_{\Diamond \mathtt{A}}$ and $t_{\Diamond \mathtt{B}}$. It thus represents the sender's epistemic state in which she considers two possibilities, namely that the receiver may take only an apple, and that the receiver may take only a pear. With this both strands of the IBR model arrive at the same unique fixed point for epistemic experts (3.1) and for non-experts (3.2):

$$R^* = \left\{ \begin{array}{lcl} m_{\Diamond \mathtt{A}} & \mapsto & t_{[\Diamond \mathtt{A}]} \\ m_{\Diamond \mathtt{B}} & \mapsto & t_{[\Diamond \mathtt{B}]} \\ m_{\Diamond (\mathtt{A} \vee \mathtt{B})} & \mapsto & t_{[\Diamond \mathtt{A}, \Diamond \mathtt{B}]} \\ m_{\Diamond (\mathtt{A} \wedge \mathtt{B})} & \mapsto & t_{[\Diamond \mathtt{AB}]} \end{array} \right\} \tag{3.1}$$

$$R^* = \left\{ \begin{array}{lcl} m_{\Diamond \mathtt{A}} & \mapsto & t_{[\Diamond \mathtt{A}]}, t_{[\Diamond \mathtt{A}, \Diamond \mathtt{AB}]}, t_{[\Diamond \mathtt{A}, \Diamond \mathtt{A} | \mathtt{B}]}, t_{[\Diamond \mathtt{A}, \Diamond \mathtt{AB}, \Diamond \mathtt{A} | \mathtt{B}]} \\ m_{\Diamond \mathtt{B}} & \mapsto & t_{[\Diamond \mathtt{B}]}, t_{[\Diamond \mathtt{B}, \Diamond \mathtt{AB}]}, t_{[\Diamond \mathtt{B}, \Diamond \mathtt{A} | \mathtt{B}]}, t_{[\Diamond \mathtt{B}, \Diamond \mathtt{AB}, \Diamond \mathtt{A} | \mathtt{B}]} \\ m_{\Diamond (\mathtt{A} \vee \mathtt{B})} & \mapsto & t_{[\Diamond \mathtt{A}, \Diamond \mathtt{B}]}, t_{[\Diamond \mathtt{A}, \Diamond \mathtt{B}, \Diamond \mathtt{AB}]}, t_{[\Diamond \mathtt{A}, \Diamond \mathtt{B}, \Diamond \mathtt{A} | \mathtt{B}]}, t_{[\Diamond \mathtt{A}, \Diamond \mathtt{B}, \Diamond \mathtt{AB}, \Diamond \mathtt{A} | \mathtt{B}]} \\ m_{\Diamond (\mathtt{A} \wedge \mathtt{B})} & \mapsto & t_{[\Diamond \mathtt{AB}]} \end{array} \right\} \tag{3.2}$$

Universal Modals. Similarly, we would hope that ignorance readings of an utterance of the universal modal statement $\Box(A \vee B)$ should fall out of the lifted model straightforwardly. But this is not quite so. When lifting the signaling game in figure 3.7, we again obtain fifteen epistemic states:

$$\begin{aligned} T \;=\; \{ & t_{[\Box \mathtt{A}]},\ t_{[\Box \mathtt{B}]},\ t_{[\Box \mathtt{A}, \Box \mathtt{B}]},\ t_{[\Box \mathtt{AB}]},\ t_{[\Box \mathtt{A}, \Box \mathtt{AB}]},\ t_{[\Box \mathtt{B}, \Box \mathtt{AB}]},\ t_{[\Box \mathtt{A}, \Box \mathtt{B}, \Box \mathtt{AB}]},\ t_{[\Box \mathtt{A} | \mathtt{B}]}, \\ & t_{[\Box \mathtt{A}, \Box \mathtt{A} | \mathtt{B}]},\ t_{[\Box \mathtt{B}, \Box \mathtt{A} | \mathtt{B}]},\ t_{[\Box \mathtt{A}, \Box \mathtt{B}, \Box \mathtt{A} | \mathtt{B}]},\ t_{[\Box \mathtt{AB}, \Box \mathtt{A} | \mathtt{B}]},\ t_{[\Box \mathtt{A}, \Box \mathtt{AB}, \Box \mathtt{A} | \mathtt{B}]}, \\ & t_{[\Box \mathtt{B}, \Box \mathtt{AB}, \Box \mathtt{A} | \mathtt{B}]},\ t_{[\Box \mathtt{A}, \Box \mathtt{B}, \Box \mathtt{AB}, \Box \mathtt{A} | \mathtt{B}]} \}. \end{aligned}$$

But here, of course, the unlifted states have to be interpreted slightly differently. So, the state $t_{[\Box \mathtt{A}, \Box \mathtt{B}]}$ is now an epistemic state of the sender where she considers only two possibilities, namely the unlifted state $t_{\Box \mathtt{A}}$ where the receiver has to take an apple (while being allowed not to take a pear), and the unlifted state $t_{\Box \mathtt{B}}$ where the receiver has to take a pear (while being allowed not to take an apple). For this model, the both the S_0-sequence as well as the R_0-sequence derive the same fixed point. For epistemic experts we get:

$$R^* = \left\{ \begin{array}{lcl} m_{\Box \mathtt{A}} & \mapsto & t_{[\Box \mathtt{A}]} \\ m_{\Box \mathtt{B}} & \mapsto & t_{[\Box \mathtt{B}]} \\ m_{\Box (\mathtt{A} \vee \mathtt{B})} & \mapsto & t_{[\Box \mathtt{A} | \mathtt{B}]} \\ m_{\Box (\mathtt{A} \wedge \mathtt{B})} & \mapsto & t_{[\Box \mathtt{AB}]} \end{array} \right\} \tag{3.3}$$

while for non-experts we get:

$$R^* = \left\{ \begin{array}{lcl} m_{\Box \mathtt{A}} & \mapsto & t_{[\Box \mathtt{A}]},\ t_{[\Box \mathtt{A}, \Box \mathtt{AB}]} \\ m_{\Box \mathtt{B}} & \mapsto & t_{[\Box \mathtt{B}]},\ t_{[\Box \mathtt{B}, \Box \mathtt{AB}]} \\ m_{\Box (\mathtt{A} \vee \mathtt{B})} & \mapsto & t_{[\Box \mathtt{A}, \Box \mathtt{B}]},\ t_{[\Box \mathtt{A}, \Box \mathtt{B}, \Box \mathtt{AB}]},\ t_{[\Box \mathtt{A} | \mathtt{B}]},\ t_{[\Box \mathtt{A}, \Box \mathtt{A} | \mathtt{B}]},\ t_{[\Box \mathtt{B}, \Box \mathtt{A} | \mathtt{B}]}, \\ & & t_{[\Box \mathtt{A}, \Box \mathtt{B}, \Box \mathtt{A} | \mathtt{B}]},\ t_{[\Box \mathtt{AB}, \mathtt{A} | \mathtt{B}]},\ t_{[\Box \mathtt{A}, \Box \mathtt{AB}, \Box \mathtt{A} | \mathtt{B}]},\ t_{[\Box \mathtt{B}, \Box \mathtt{AB}, \Box \mathtt{A} | \mathtt{B}]}, \\ & & t_{[\Box \mathtt{A}, \Box \mathtt{B}, \Box \mathtt{AB}, \Box \mathtt{A} | \mathtt{B}]} \\ m_{\Box (\mathtt{A} \wedge \mathtt{B})} & \mapsto & t_{[\Box \mathtt{AB}]} \end{array} \right\} \quad (3.4)$$

These predictions are not correct. For expert senders, the target form $m_{\Box(\mathtt{A} \vee \mathtt{B})}$ should be interpreted as $t_{[\Box \mathtt{A}, \Box \mathtt{B}]}$ instead, because in a context where the sender is not an absolute authority, a sentence like (46a) should implicate both (46b) and (46c).

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

b. $\rightsquigarrow$ The speaker considers it possible that the hearer *must* take an apple (a pear).

c. $\rightsquigarrow$ The speaker considers it possible that the hearer *need not* take an apple (a pear).

Unfortunately, the IBR model only predicts (46c) and not (46b). As a matter of fact, the IBR model predicts the expert sender to be too much of an expert. This problem did not arise under existential modals because there the alternative forms $m_{\Diamond \mathtt{A}}$ and $m_{\Diamond \mathtt{B}}$ were true in the corresponding state $t_{\Diamond \mathtt{A} | \mathtt{B}}$. Under universal modals, however, the only message that is true in $t_{\Box \mathtt{A} | \mathtt{B}}$ is the target message $m_{\Box(\mathtt{A} \vee \mathtt{B})}$. This way, when the IBR model looks for the most informed sender state where $m_{\Box(\mathtt{A} \vee \mathtt{B})}$ is true, it finds a too specific state $t_{[\Box \mathtt{A} | \mathtt{B}]}$, instead of the intuitively correct $t_{[\Box \mathtt{A}, \Box \mathtt{B}]}$.

Reprehensibility. This problem could perhaps be solved by arguing for a different set of alternatives to $m_{\Box(\mathtt{A} \vee \mathtt{B})}$. Another, to my mind more interesting, strategy is to assume an adequate order on the set of states. This is what many alternative accounts of FC- and ignorance implicatures rely on, and it is already astonishing enough that the IBR model derives FC-readings for both existential and universal modals, as well as ignorance readings for existential modals, without such extra ordering information. A plain but appealing first shot at characterizing minimality of a deontic state based on a set of relevant propositions P is to say that a state t is MORE RESTRICTED IN ITS PERMISSIONS than another state t' iff t' makes more sentences of the form $\Diamond p$ for $p \in P$ true

than t does (cf. van Fraassen 1973; Kratzer 1981; Lewis 1981). Analogously, a state t is MORE RESTRICTED IN ITS OBLIGATIONS than another state t' iff t' makes more sentences of the form $\Box p$ for $p \in P$ true than t does. In a signaling game context, the minimality of models could be translated into an assumption about the prior probabilities of states: more minimal models are *a priori* more likely because these are the stereotypical interpretations that first spring to mind. For the signaling game in figure 3.7, this latter notion would induce an ordering on prior probabilities as follows:

$$\Pr(t_{\Box \mathtt{A}|\mathtt{B}}) \;>\; \Pr(t_{\Box \mathtt{A}}) = \Pr(t_{\Box \mathtt{B}}) \;>\; \Pr(t_{\Box \mathtt{AB}})$$

If we allow this ordering information —which, by the way, does not disturb predictions for the unlifted game— to take precedence over the ordering information on epistemic states from speaker expertise, the model predicts the intuitively correct outcome also for ignorance implicatures under universal modals. Whether this is a generally and conceptually satisfactory solution, I will have to leave for another occasion.

3.3.3 Simplification of Disjunctive Antecedents

Although we will come back in detail to questions concerning the interpretation of conditionals in chapter 5 , I would like to round off the discussion of FC-inferences by a brief look at conditionals with a disjunctive antecedent like in (47a).

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

b. $\rightsquigarrow$ If you eat an apple, you will feel better. $A > C$

c. $\rightsquigarrow$ If you eat a pear, you will feel better. $B > C$

(48) a. If you'd eaten an apple or a pear, you'd feel better. $(A \vee B) > C$

b. $\rightsquigarrow$ If you'd eaten an apple, you'd feel better. $A > C$

c. $\rightsquigarrow$ If you'd eaten a pear, you'd feel better. $B > C$

Intuitively, the indicative (47a) seems to convey both (47b) and (47c), and similarly the counterfactual (48a) seems to convey both (48b) and (48c). 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 necessarily make SDA valid.[23] This 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.

Still, there are good arguments not to want SDA to be a semantically valid inference pattern. Warmbrōd (1981) gives one 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 (49a) to (49b) are generally valid, which intuitively should not be the case.[24]

(49) a. If you eat an apple, you will feel better. $A > C$

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

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

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

b. $\not\rightarrow$ If John had taken a pear, he would have taken an apple.

If SDA was semantically valid then (50a) would imply (50b), but this is of course nonsense. Together, this suggests loosely that SDA should perhaps be thought of as a pragmatic inference on top of a standard semantics.

A pragmatic account is moreover also made plausible by the observation that SDA is structurally very similar to FC-readings (see Klinedinst 2006; van Rooij 2006*a*). Asher and Bonevac (2005) even analyze permission statements of the form "you may do A" as, roughly, a conditional statement "if you do A, it is okay." This is also very plausible given the fact that an English question like

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

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

23. I will not enlarge on semantic theories of conditionals here. Readers unfamiliar with this topic may want to skip ahead and consult section 5.1.

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

(52) 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.'

A final parallel between SDA and FC is the observation that we can force epistemic ignorance readings also for conditionals with disjunctive antecedents (see Klinedinst 2006):

(53) a. If you eat an apple or a pear, you will feel better, but I don't know which. $(A \vee B) > C$

b. $\rightsquigarrow$ The speaker considers $A > C$ possible, but not necessary.

c. $\rightsquigarrow$ The speaker considers $B > C$ possible, but not necessary.

In a context like (53a) that marks the speaker's epistemic uncertainty we do not derive from $(A \vee B) > C$ that the speaker *knows* that $A > C$ and $B > C$ are both true, as full-fledged SDA would have it. Rather, if we take the speaker to be maximally knowledgeable despite her expressed uncertainty, we only infer that the speaker considers exactly one of the sentence $A > C$ and $B > C$ true, but not both.

For these reasons, we should try and see whether SDA can be derived as a pragmatic inference similar to FC-readings in the IBR model. It seems that the exact same approach that we used for FC-readings and ignorance readings above should apply also for SDA and ignorance readings such as in (53). In particular, it may be suspected that SDA as in (47) and (48) can be explained as a general pragmatic inference associated with conditionals $(A \vee B) > C$ in standard unlifted signaling games. The ignorance readings in a context which forces us to assume speaker uncertainty, like in (53), should also be explicable, as before, in terms of lifted signaling games. If we could thus explain SDA as an inference by iterated pragmatic reasoning this would also rebut the claim of localists Levinson (2000) and Chierchia et al. (2008) that sequences like (54) force Gricean reasoning to penetrate into syntax so that an embedded implicature is calculated under the scope of the antecedent operator.

(54) If you take an apple or a pear, that's fine. But if you take both, that's not okay.

CONTEXT MODEL. Following our general principles for construction of context models, we should start with a suitable set of expression alternatives to

the target expression $m_{(\mathtt{A}\vee\mathtt{B})>\mathtt{C}}$. In line with the previous treatment of disjunction it is safe to assume three further alternative expressions, namely $m_{\mathtt{A}>\mathtt{C}}$, $m_{\mathtt{B}>\mathtt{C}}$, and $m_{(\mathtt{A}\wedge\mathtt{B})>\mathtt{C}}$ with the obvious intended meanings. The question then is which semantics we should adopt for conditional sentences. Let me defer more in-depth discussion of conditional semantics to chapter 5, and confine myself here to just stating the abstract semantic scheme that I will endorse in this thesis for both indicatives and counterfactuals.

Let each possible world w be associated with a MODAL STRUCTURE $\langle R_w, \preceq_w \rangle$ that is suitable for interpreting the conditional that we are interested in. Generally, R_w is a set of possible worlds and $\preceq_w$ is a well-founded ordering on R_w. Many reasonable constraints on the nature of this ordering could be given to instantiate certain influential theories of conditionals (think of: Stalnaker 1968; Lewis 1973; Kratzer 1981; Lewis 1981; Veltman 1985). For the present pragmatic purpose we should remain noncommittal and not take on *any* particular constraints on modal structures. We then simply define

$$\mathrm{Min}(R_w, \preceq_w, A) = \left\{ v \in R_w \cap A \mid \neg\exists v' \in R_w \cap A \ : \ v' \prec_w v \right\}$$

and say that an indicative or counterfactual conditional

$$A > C \text{ is true in } w \quad \text{iff} \quad \mathrm{Min}(R_w, \preceq_w, A) \subseteq C.$$

To derive the states of our signaling game, we should then look at the eight conjunctive combinations of alternative forms in the following table and ask which of these combinations are consistent:

	$m_{\mathtt{A}>\mathtt{C}}$	$m_{\mathtt{B}>\mathtt{C}}$	$m_{(\mathtt{A}\wedge\mathtt{B})>\mathtt{C}}$
t_1	$\surd$	$\surd$	$\surd$
t_2	$\surd$	$\surd$	$-$
t_3	$\surd$	$-$	$\surd$
t_4	$\surd$	$-$	$-$
t_5	$-$	$\surd$	$\surd$
t_6	$-$	$\surd$	$-$
t_7	$-$	$-$	$\surd$
t_8	$-$	$-$	$-$

As $(A \vee B) > C$ implies $(A > C) \vee (B > C)$ under the assumed general semantics, states t_7 and t_8 are inconsistent. All other combinations are possible and non-redundant, and so we end up with six possible states in the context model given in figure 3.8.

	$\Pr(t)$	t_1	t_2	t_3	t_4	t_5	t_6
t_1	$1/6$	1,1	0,0	0,0	0,0	0,0	0,0
t_2	$1/6$	0,0	1,1	0,0	0,0	0,0	0,0
t_3	$1/6$	0,0	0,0	1,1	0,0	0,0	0,0
t_4	$1/6$	0,0	0,0	0,0	1,1	0,0	0,0
t_5	$1/6$	0,0	0,0	0,0	0,0	1,1	0,0
t_6	$1/6$	0,0	0,0	0,0	0,0	0,0	1,1

	$m_{\mathtt{A}>\mathtt{C}}$	$m_{\mathtt{B}>\mathtt{C}}$	$m_{(\mathtt{A}\wedge\mathtt{B})>\mathtt{C}}$	$m_{(\mathtt{A}\vee\mathtt{B})>\mathtt{C}}$
t_1	$\surd$	$\surd$	$\surd$	$\surd$
t_2	$\surd$	$\surd$	$-$	$\surd$
t_3	$\surd$	$-$	$\surd$	$\surd$
t_4	$\surd$	$-$	$-$	$\surd$
t_5	$-$	$\surd$	$\surd$	$\surd$
t_6	$-$	$\surd$	$-$	$\surd$

Figure 3.8: Unlifted context model for SDA

PREDICTIONS. The IBR model predicts a slightly different fixed point for each IBR sequence. The interpretation of the target message, however, is the exact same in both fixed points. Let us first look at the predictions for unlifted games. For the R_0-sequence, we obtain

$$R^* = \left\{ \begin{array}{lcl} m_{\mathtt{A}>\mathtt{C}} & \mapsto & t_4 \\ m_{\mathtt{B}>\mathtt{C}} & \mapsto & t_6 \\ m_{(\mathtt{A}\wedge\mathtt{B})>\mathtt{C}} & \mapsto & t_1, t_3, t_5 \\ m_{(\mathtt{A}\vee\mathtt{B})>\mathtt{C}} & \mapsto & t_2 \end{array} \right\}$$

as fixed-point interpretation behavior. For the S_0-sequence, on the other hand, we obtain the same, except that

$$R^*(m_{(\mathtt{A}\wedge\mathtt{B})>\mathtt{C}}) = \{t_3, t_5\}\,.$$

Still, the interpretation of our target message $m_{(\mathtt{A}\vee\mathtt{B})>\mathtt{C}}$ is exactly as it should be in accordance with SDA. The state t_2 is indeed one where both $A > C$ and $B > C$ are true but where $(A \wedge B) > C$ is false.[25] I am content with this result

25. As before for the derivation of FC-readings, it may be contestable that $(A \wedge B) > C$ is derived to be false. As before this inference may require some emphatic stress or more contextual relevance of the conjunction alternative. *Without* the conjunctive alternative the

but admit that the diverging interpretation of $m_{(\mathtt{A}\wedge\mathtt{B})>\mathtt{C}}$ between IBR sequences is a minor oddity of the model.[26]

As for the lifted models, the situation is similar. Assuming epistemic experts, the R_0-sequence reaches a fixed point in the interpretation strategy

$$R^* = \left\{ \begin{array}{lcl} m_{\mathtt{A}>\mathtt{C}} & \mapsto & t_{[4]} \\ m_{\mathtt{B}>\mathtt{C}} & \mapsto & t_{[6]} \\ m_{(\mathtt{A}\wedge\mathtt{B})>\mathtt{C}} & \mapsto & t_{[1]}, t_{[3]}, t_{[5]} \\ m_{(\mathtt{A}\vee\mathtt{B})>\mathtt{C}} & \mapsto & t_{[4,5]}, t_{[3,6]}, t_{[4,6]} \end{array} \right\}$$

while the interpretation fixed point of the S_0-sequence differs only in the interpretation of the conjunctive alternative:

$$R^*(m_{(\mathtt{A}\wedge\mathtt{B})>\mathtt{C}}) = \left\{ t_{[3,5]} \right\}.$$

We derive the intuitively attested epistemic readings for a case like (53): the sender is taken to believe that exactly one out of $A > C$ and $B > C$ is true, while being uncertain which one that is.

For completeness, let me also give the predictions of the model for the inexpert case, at least for our target message. Predictions are the same here for both IBR sequences:

$$R^*(m_{(\mathtt{A}\vee\mathtt{B})>\mathtt{C}}) = \left\{ \begin{array}{l} t_{[2,3,5]},\ t_{[1,2,3,5]},\ t_{[4,5]},\ t_{[1,4,5]},\ t_{[2,4,5]},\ t_{[1,2,4,5]}, \\ t_{[3,4,5]},\ t_{[1,3,4,5]},\ t_{[2,3,4,5]},\ t_{[1,2,3,4,5]},\ t_{[3,6]},\ t_{[1,3,6]}, \\ t_{[2,3,6]},\ t_{[1,2,3,6]},\ t_{[4,6]},\ t_{[1,4,6]},\ t_{[2,4,6]},\ t_{[1,2,4,6]}, \\ t_{[3,4,6]},\ t_{[1,3,4,6]},\ t_{[2,3,4,6]},\ t_{[1,2,3,4,6]}, \\ t_{[3,5,6]},\ t_{[1,3,5,6]},\ t_{[2,3,5,6]},\ t_{[1,2,3,5,6]}, \\ t_{[4,5,6]},\ t_{[1,4,5,6]},\ t_{[2,4,5,6]},\ t_{[1,2,4,5,6]},\ t_{[3,4,5,6]}, \\ t_{[1,3,4,5,6]},\ t_{[2,3,4,5,6]},\ t_{[1,2,3,4,5,6]} \end{array} \right\}.$$

Although unwieldy, these results are intuitive, as is easy to check. We predict the inference attested in (53) that the sender considers both $A > C$ and $B > C$ possible but does not have enough information to believe any one true. Unlike for epistemic experts we now no longer obtain that the sender believes

signaling game model is fairly trivial and predictions are unremarkable: SDA is derived from the unlifted model with the same fixed point for both sequences, and the lifted models also derive the obvious intuitive results.

26. Still this is not badly worrisome, because interpretation of the form $m_{(\mathtt{A}\wedge\mathtt{B})>\mathtt{C}}$ would trigger a different context representation, and so, strictly speaking, we always have to worry only about the interpretation of the target message.

that only one of these conditionals is true; there are epistemic states associated with our target message that contain t_1 or t_2, i.e., there are states in the interpretation of $m_{(\mathtt{A}\vee\mathtt{B})>\mathtt{C}}$ where the sender considers it possible that $A > C$ and $B > C$ are both true at the same time.

Summary. The IBR model offers a parallel solution for free-choice readings of disjunctions under modals, and also for the related SDA-inferences if we consult unlifted signaling games. Lifted game models naturally account for the epistemic readings associated with both types of constructions.

3.4 Games at the Semantics-Pragmatics Interface

As the true philosopher that Grice was, he managed to inspire by raising the right questions rather than by providing fully resolving answers. To fill in the details of the Gricean programme was left to a community of philosophers and linguists, and more recently also psycholinguists. The issues debated in connection with Grice's notion of implicature, and linguistic and speaker meaning are still very much alive. After having detailed the IBR model and shown some of its applications, it is time to place game theoretic pragmatics in its current variety on the map by showing its position in some of the relevant controversies about the interface between semantics and pragmatics.

Global or Local. A first issue that should be addressed because it has recently been vividly debated is whether conversational implicatures are to be computed globally or locally. To see what is at stake, take again the case (54), repeated here, that we have just looked at in section 3.3.3.

(54) If you take an apple or a pear, that's fine. But if you take both, that's not okay.

Intuitively speaking, the scalar inference associated with "or" seems to take scope under the meaning of "if" and that may suggest that whenever scalar items fall into the scope of other operators, the scalar inference should be computed locally within the narrow scoping. A strong local view would therefore require that implicatures be computed as part of syntax (cf. Levinson 2000; Chierchia 2004; Fox 2007; Chierchia et al. 2008). In contrast, a scalar inference is computed globally if it is derived by comparing alternatives to a target

scalar item in the full linguistic context of its occurrence, e.g., by comparing the whole conditional in (54) to other alternative conditionals without a disjunctive antecedent.

As many of the previous examples showed, the IBR model clearly takes and supports a global approach to scalar implicature calculation (cf. van Rooij and Schulz 2004; Schulz and van Rooij 2006; Russell 2006). A major contribution of in particular this chapter is the proof that many allegedly local scalar inferences can be accounted for, especially if iteration of optimality considerations is taken into account. Indeed, as far as scalar inferences are concerned, I fully endorse the view of Geurts (2009) who argues that only very few marked cases seem to resist a global treatment.

Pragmatic Intrusion and the Gazdarian Picture. But although I would preferably apply the IBR model as a globalist reasoning scheme when it comes to *scalar* inferences, that does not mean that the IBR model is actually committed to a rigid modular architecture in which *all* pragmatic inference takes place based on fully spelled out truth-conditional semantics. To appreciate this point fully, let us briefly take a step back and recapitulate some of the basic ideas about the relation between semantics and pragmatics.

Grice himself had suggested that conversational implicatures should be derivable from "what was said" together with the Cooperative Principle and the Maxims of Conversation. But there is still an ongoing debate about a clear demarcation between semantic meaning and "what was said" on the one hand, and conversational implicatures and "what was meant" on the other. On one end of the (multidimensional) spectrum, we find positions like Gerald Gazdar's who holds that utterance meaning is computed globally and modularly: according to Gazdar, Gricean inference takes semantic meaning, which is truth-conditional meaning unmediated by pragmatic processes, as a starting point (Gazdar 1979). Opposed to this strictly modular picture, others have acknowledged the role of Gricean inferences already in establishing the truth-conditional meaning of an utterance, such as for instance in *expanding* (55a) or *completing* (55b) a proposition (see Carston 1988; Recanati 1989; Bach 1994; Levinson 2000; Recanati 2004).

(55) a. You are are not going to die.
$\rightsquigarrow$ You are not going to die *from this*.

b. Keisuke was too late.
$\rightsquigarrow$ Keisuke was too late *for pie*.

Contrary to superficial impression, the IBR model is not committed to a strict Gazdarian conception, but is entirely compatible with the idea that certain pragmatic inferences feed the specification of sentence meaning, based on which the IBR model may kick in and do its work. Although I have assumed that messages in the game model have traditional truth-conditional semantics, this is —as I have already mentioned in section 3.1— not at all necessary. The IBR model could equally well deal with fairly weak conceptions of semantic meaning (see Borg 2004; Recanati 2004; Cappelen and Lepore 2005), as long as we may assume that (the interpreter assumes that) a semantic meaning uniquely exists that is shared and commonly accessible. The IBR model thus seems incompatible with only the most extreme 'anything goes' theories of conventional meaning (such as found, for instance, in a strong reading of Davidson 1986).

GENERALIZED OR PARTICULARIZED INFERENCES. Grice distinguished GENERALIZED CONVERSATIONAL IMPLICATURES that seem to occur for some given lexical material with a certain predictable regularity from PARTICULARIZED CONVERSATIONAL IMPLICATURES that arise for seemingly arbitrary lexical material and only under special contextual constellations. The inferences associated with scalar items like "some" or "possibly" are prime examples of generalized implicatures. But scalar inferences also occur for more *ad hoc* comparisons between possible utterances: if we went shopping together and you know that we bought Gouda and Emmentaler cheese, then if I say

(56) I ate the Emmentaler.

you may take this to mean that I did not eat the Gouda. But clearly an out-of-the-blue utterance of (56) would not trigger this inference. This is then a clear example of a particularized implicature.

On the face of it, the present game theoretic approach treats all pragmatic inferences as reasoning about language use in a given context and consequently mainly accounts for particularized implicatures that arise from particular hearer beliefs about the concrete utterance context. However, by reference to interpretation games as representations of generic contexts of sentence interpretation, the present approach nonetheless also covers generalized implicatures as those inferences associated with utterances of sentences in an out-of-the-blue context. A similar contextualist view underlies not only game theoretic approaches (Benz and van Rooij (2007) are very outspoken on this issue) but also relevance theory (Sperber and Wilson 1995; Carston 1998) and

Neo-Gricean approaches with a clear affinity towards rational choice models of utterance contexts (see van Rooij and Schulz 2004, 2006; Schulz and van Rooij 2006).

This contextualist view is opposed to the idea that generalized implicatures have a special *default status*, a theory that is supported by, for instance, Levinson (2000) and Chierchia (2004). But there are good empirical arguments against the idea that generalized implicatures are special and/or computed as a default (see Noveck and Sperber 2004; Katsos 2008*b*, for overview on experimental approaches to pragmatics). Experimental data offers evidence that only if a scalar implicature arises in context its computation does take time (Noveck and Posada 2003; Bott and Noveck 2004). This clearly speaks against a default approach which would predict the reverse pattern. Other studies similarly stress the importance of context in computation of implicatures (see Breheny et al. 2006; Katsos 2008*b*). Specifically, there is compelling evidence that whether a scalar inference arises or not crucially hinges on the contextual question under discussion (see Zondervan 2006). Finally, both young language learners as well as adults seem to reason just as proficiently, if not even better, with contextualized ad hoc alternatives of the variety in (56) as with generalized lexical alternatives (see Katsos and Bishop 2009). All of this taken together supports the view that implicatures are contextualized, in line with the present approach.

On top of empirical arguments, there are also conceptual arguments in favor of the contextualist position. The main advantage of the present game theoretic approach in this respect is that we have very rich and explicit context models. Obviously, games can model very fine distinctions both in the beliefs of interlocutors as well as in the preferences of individual agents. This can be relevant for linguistic interpretation in diverse ways. For instance, under normal circumstances the answer to a question like in (57) is interpreted exhaustively as implicating that Bill did not come, but the answer to a question like that in (58) is not.

(57) a. Who, of John, Bill and Mary, came to the party?

b. John and Mary did.

c. $\rightsquigarrow$ Bill did not.

(58) a. Where can I get an Italian newspaper?

b. At the reception.

c. $\not\rightsquigarrow$ Not at the airport.

The reason for this difference in interpretation of answers intuitively lies in the relevance that certain information has for the questioner based on a practical decision he faces (see van Rooij 2003*b*). To account for the structural commonalities and differences of cases (57) and (58), models that represent an agent's individual preferences in a goal-oriented setting are advantageous if not necessary. A detailed representation of individual preferences thus pins down what exactly is *relevant* for the conversationalists, independent of lexicalized scales (cf. Benz 2007). The crucial point is that rational choice models not only always incorporate a notion of relevance, but also *reduce* it in a natural way to individual preferences.[27]

Moreover, games as context models not only include the preferences of *single* agents, but crucially those of *all* discourse participants. This lets us model different levels of partial alignment or divergence of preferences of multiple agents. Grice's assumption of cooperation in conversation is easily integrated as a special case, but it is clear that the representative power of games provides much more generality. Game models let us represent arbitrary constellations of partially cooperative, partially adversary discourses. Predictions are not confined to cooperation only —as in traditional Gricean approaches— or to argumentation only —as for instance in the work of Ducrot (1973), Anscombre and Ducrot (1983) and Merin (1999)— and this makes GTP of the current variety much more general and systematically applicable than other approaches (see also van Rooij 2004*a*; Benz 2006; Franke et al. to appear, as well as section 2.5).

In sum, the present GTP approach is a very flexible, but nonetheless rigorous, contextualist approach to pragmatic inferences. Both on-the-fly context-dependent reasoning as well as sentence interpretation in generic contexts can be accommodated in a uniform theory that is backed up by both empirical evidence as well as conceptual considerations.

NATURALISTIC OR NORMATIVE. A final issue, squarely related to the distinction between default and contextual accounts, is whether the present approach understands itself as a *naturalistic description* of actual reasoning about language or rather as a *normative prescription* of how we ought to reason. At first glance, the IBR model has elements of both, and which interpretation is

27. A further advantage of this is that a preference for informativity, as postulated in Grice's Maxim of Quantity and upheld by the Neo-Griceans, falls out as a *special case* in preference-based approaches, just as it should. This argument is presented by Bernardo (1979) in the abstract, and by van Rooij (2004*c*) in the context of natural language interpretation.

most plausible may seem to depend on the intended application.

My preferred view on the matter certainly showed in the way I have applied the model so far. I take it that the IBR model aims to explain actual linguistic *competence* and not necessarily performance with all conceivable interferences factored in. Still, I would like to think of the IBR model as a descriptive, not a prescriptive approach. This is because I tend to think of the model as an account of *idealized* reasoning behavior, rather than as a full-fledged performance model, despite the fact that the IBR model as such includes certain natural restrictions on reasoning competence, such as the focality of conventional meaning or a tendency towards unbiased belief formation. The IBR model thus seeks to balance a formally rigorous and predictive approach in the vein of the Neo-Griceans with the cognitive realism advocated by relevance theorists: it tries to explain pragmatic competence as rational inference given further psychologically plausible assumptions about the cognitive architecture of reasoners.

This is also to say that I vehemently reject any commitment to the absurd notion that every time a proficient speaker of English grasps, say, a free choice inference, she has gone *consciously* through exactly the calculation the IBR model offers for this inference. In particular, although in derivations of implicatures I have mostly consulted the limit prediction of the IBR model, I am only committed to the idea that proficient language users are *in principle* able to carry out such intricate higher-order theory of mind reasoning steps, not that they actually perform these as a conscious reasoning process every time anew. Empirical research suggests that in certain domains and under certain conditions taking other people's perspective into account may happen immediately and automatically (cf. Hanna et al. 2003; Heller et al. 2008), but such processes also seem costly (Keysar et al. 2003). So the IBR model may better be conceived of as a model of perhaps subconscious optimization in production and interpretation that requires competence of higher-order theory of mind reasoning, but not necessarily repeated execution thereof once a piece of pragmatic competence is mastered. The next chapter also furthermore addresses issues of perspective-taking in language use and moreover the acquisition of pragmatic competencies in language learners whose ToM capabilities might not yet match adult competence.

Chapter 4

Perspective, Optimality & Acquisition

Chapter Contents

4.1 · Optimality Theory in Pragmatics · 182

4.2 · BiOT and Game Theory · 190

4.3 · An Epistemic Interpretation of Optimality · 196

4.4 · Scalar Implicatures in Language Acquisition · 213

This chapter compares the IBR model with bidirectional optimality theory (BIOT), an alternative formal framework for Gricean pragmatics, as introduced by Blutner (1998, 2000) (see Blutner and Zeevat 2008, for a recent overview). The chapter is structured as follows. Section 4.1 will first briefly introduce optimality theory (OT) in general, and then zoom in on the use of BIOT in the context of Gricean pragmatics. Subsequently we will explore ways of comparing optimality theoretic with game theoretic pragmatics. Section 4.2 reviews critically a previous characterization of OT in terms of strategic games. Section 4.3 draws a different picture of the connection between OT-pragmatics and game theory: I argue that OT-pragmatics should be linked to signaling games, and attempt an epistemic characterization of optimality notions in terms of particular restrictions on the belief formation strategies of IBR reasoners.

4.1 Optimality Theory in Pragmatics

Optimality theory has its origin in phonology (Prince and Smolensky 1997), but has been readily applied to other linguistic subdisciplines such as syntax, semantics (Hendriks and de Hoop 2001), and pragmatics (c.f. the contributions in Blutner and Zeevat 2004).

4.1.1 OT-Systems

Abstractly speaking, OT is a model of how input and output representations are associated with each other based on a set of ranked, violable constraints that express relative preferences for input-output matching. More concretely, an OT-SYSTEM for a set M of (input) forms and a set T of (output) meanings[1] is just a pair $\langle \text{Gen}, \succeq \rangle$ consisting of a GENERATOR $\text{Gen} \subseteq M \times T$ that gives us the initially possible form-meaning pairs and an ordering $\succeq$ on elements of Gen. For Gen and other sets O of form-meaning pairs, write:

$$\begin{aligned} O(t) &= \{m \in M \mid \langle m, t \rangle \in O\} \\ O(m) &= \{t \in T \mid \langle m, t \rangle \in O\} \end{aligned}$$

1. For the pragmatic applications that we are interested in here, the inputs of an OT-system are (representations of) linguistic forms and the outputs are (representations of) meanings. In anticipation of a comparison between game theoretic and optimality theoretic pragmatics, I will write M for the set of forms and T for the set of meanings. For the time being, these are just variables. We will come back later to the issue of identifying forms and meanings in OT with messages and states in a game theoretic setting.

and let $\succeq_m$ and $\succeq_t$ be the orderings induced by $\succeq$ on the sets $\text{Gen}(m)$ and $\text{Gen}(t)$ respectively:

$$\begin{aligned} t \succeq_m t' \quad &\text{iff} \quad \langle m, t\rangle \succeq \langle m, t'\rangle \\ m \succeq_t m' \quad &\text{iff} \quad \langle m, t\rangle \succeq \langle m', t\rangle\,. \end{aligned}$$

I will assume for simplicity that every OT-system is such that $\succeq_m$ and $\succeq_t$ are well-founded, linear orders on $\text{Gen}(m)$ and $\text{Gen}(t)$ for all m and t.[2]

The ordering of an OT-system measures how well the elements of the generator satisfy certain standards of grammaticality, normality, efficiency, or whatever might be at stake for a particular application. For instance, $m \succeq_t m'$ would mean that m is (somehow) a better form for meaning t then m' is. What exactly the ordering measures may be left unspecified if we just want to assess the general architecture of OT-systems. It might also be defined directly in terms of properties of the elements of the generator. But for most applications of OT, the ordering is actually derived from a (finite) set Con of violable constraints that are ranked with respect to importance by an ordering $\gg$. Constraints in Con compare elements in Gen with respect to other elements in Gen according to some criterion of preferred input-output matching. Abstractly, for a given set Gen each constraint in Con is just a mapping from Gen to a natural number, possibly zero, specifying the number of times and/or the magnitude that a given form-meaning pair violates the constraint in question. There are two kinds of constraints:

(i) MARKEDNESS CONSTRAINTS compare either only the input dimension or only the output dimension;

(ii) FAITHFULNESS CONSTRAINTS compare input-output pairs to each other based on how well each pair's input associates with its output.

An easy example of a markedness constraint is the processing cost of a form: if a form m' is more costly to process than a form m, then this can be expressed in a markedness constraint C for which $C(\langle m', \cdot\rangle) > C(\langle m, \cdot\rangle)$ irrespective of the meaning component in the to-be-compared pairs. An easy example for a faithfulness constraint is Gricean Quality: a pair $\langle m, t\rangle$ would violate this constraint just in case the meaning t is incompatible with the semantic meaning of the form m; so for this comparison we crucially need to refer to both dimensions of the form-meaning pair.

2. Let $\succeq_m$ be well-founded on $\text{Gen}(m)$ if for all subsets $X \subseteq \text{Gen}(m)$ there is at least one $\succeq$-maximal element in X. Analogously for $\succeq_t$ and $\text{Gen}(t)$.

The ordering $\gg$ on the set Con represents the importance of the constraints relative to each other. In the simplest case $\gg$ is a linear order and the set of constraints $C_1 \gg C_2 \gg \ldots \gg C_n$ can be enumerated starting with the most important and ending with the least important constraint. For a linearly ordered set of constraints, we could think of Con as a mapping of each element of Gen to an n-tuple $\text{Con}(\langle m, t\rangle) = \langle c_1, c_2, \ldots, c_n\rangle$ of natural numbers, where each c_i is just the number $C_i(\langle m, t\rangle)$, i.e., the number of times and/or the magnitude that the input-output pair $\langle m, t\rangle$ violates the constraint C_i.

Finally, the ordering $\succeq$ of the OT-system is derived from the number or severity of the violations of the ranked constraints. For a linearly ordered set of constraints we obtain an ordering with the desired properties from the following definition. Let $g, g' \in \text{Gen}$:

$$\begin{aligned} g \succeq g' \quad \text{iff} \quad & \exists i\, \forall j < i :\ C_j(g) = C_j(g') \text{ and } C_i(g) < C_i(g') \\ & \text{or } \forall i :\ C_i(g) = C_i(g'). \end{aligned}$$

To give life to an abstract OT-system for applications we need to define the inputs and outputs and, most importantly, the ordering on the generator in some reasonable way. In the present context we are particularly interested in PRAGMATIC OT-SYSTEMS in which form-meaning pairs are evaluated by an ordering that formally captures how well —relative to others— a primitive form-meaning pair satisfies certain basic pragmatic principles.

4.1.2 Uni- and Bidirectional Optimality

Based on an ordering $\succeq$ that is either derived from Con or otherwise defined as an ordering on Gen, an OT-system can specify the preferred input-output associations in several ways. Since $\succeq$ is an ordering on a set of input-output pairs, we can either take a production perspective and ask which output is best when we fix the input dimension, or we can take a comprehension perspective and ask which input is best when we fix the output dimension. The former production perspective is taken by OT-syntax, the latter comprehension perspective is taken by OT-semantics. Abstractly, we can define the set of UNIDIRECTIONALLY OPTIMAL PAIRS as follows:

$$\begin{aligned} \text{OT}_{\text{syn}} &= \left\{\langle m, t\rangle \in \text{Gen} \mid \neg\exists t' :\ \langle m, t'\rangle \in \text{Gen} \wedge t' \succ_m t\right\} \\ \text{OT}_{\text{sem}} &= \left\{\langle m, t\rangle \in \text{Gen} \mid \neg\exists m' :\ \langle m', t\rangle \in \text{Gen} \wedge m' \succ_t m\right\} \end{aligned}$$

Optimization along both dimensions at the same time is also possible, of course. This is BIDIRECTIONAL OPTIMALITY and it comes in two varieties, a

strong notion and a weak notion (Blutner 1998, 2000). We say that an input-output pair is STRONGLY OPTIMAL iff it is unidirectionally optimal for both production and comprehension: let

$$\text{BIOT}_{\text{str}} = \text{OT}_{\text{syn}} \cap \text{OT}_{\text{sem}}$$

be the set of all strongly optimal pairs. The definition of weak optimality is a bit more intricate. Adopting Jäger's reformulation of Blutner's original definition (Jäger 2002), we say that a pair $\langle m, t \rangle$ is WEAKLY OPTIMAL iff

(i) there is no weakly optimal $\langle m, t' \rangle$ such that $t' \succ_m t$; and

(ii) there is no weakly optimal $\langle m', t \rangle$ such that $m' \succ_t m$;

and we denote the set of all weakly optimal pairs with $\text{BIOT}_{\text{weak}}$. It is obvious that all strongly optimal pairs are also weakly optimal, but it may be the case that there are weakly optimal pairs which are not strongly optimal.

Unfortunately, the recursive definition of weak optimality is somewhat difficult to apply. In practice, therefore, most often weakly optimal pairs are computed via a manageable algorithm which iteratively computes optimal pairs.[3] The BIOT-algorithm given in figure 4.1 iteratively computes three disjoint sets of form-meaning pairs:

(i) the set Pool_n of form-meaning pairs still in competition for optimality after n rounds of iteration;

(ii) the set Opt_n of form-meaning pairs that have been identified as optimal after round n;

(iii) the set Blo_n of form-meaning pairs that are blocked by an optimal pair and therefore removed from the pool.

Initially, Pool_0 is the set Gen and there are no optimal or blocked forms. The algorithm then iteratively computes optimal pairs based on a comparison of forms left in the pool and removes optimal and blocked pairs from the pool until every form-meaning pair is removed from the pool as either optimal or blocked. We could think of the pool at round n as a reduced OT-system. The BIOT-algorithm thus repeatedly checks for strong optimality in ever more reduced OT-systems.

Let me briefly mention two obvious but relevant properties of the BIOT-algorithm: firstly, $\text{Opt}_1 = \text{BIOT}_{\text{str}}$, and secondly, $\text{Opt}_n \subseteq \text{Opt}_{n+1}$ and $\text{Blo}_n \subseteq$

3. This algorithm is widely used in practice and goes back to Jäger (2002).

$\text{Pool}_0 \leftarrow \text{Gen}$
$\text{Opt}_0 \leftarrow \emptyset$
$\text{Blo}_0 \leftarrow \emptyset$
$n \leftarrow 0$
while $\text{Pool}_n \neq \emptyset$ **do**
  $\text{Opt}_{n+1} \leftarrow \text{Opt}_n \cup \{\langle m, t\rangle \in \text{Pool}_n \mid$
    $\neg\exists \langle m', t\rangle \in \text{Pool}_n \ \langle m', t\rangle > \langle m, t\rangle \wedge$
    $\neg\exists \langle m, t'\rangle \in \text{Pool}_n \ \langle m, t'\rangle > \langle m, t\rangle\}$
  $\text{Blo}_{n+1} \leftarrow \text{Blo}_n \cup \{\langle m, t\rangle \in \text{Pool}_n \mid$
    $\exists \langle m', t\rangle \in \text{Opt}_{n+1} \ \langle m', t\rangle > \langle m, t\rangle \vee$
    $\exists \langle m, t'\rangle \in \text{Opt}_{n+1} \ \langle m, t'\rangle > \langle m, t\rangle\}$
  $\text{Pool}_{n+1} \leftarrow \text{Pool}_0 \setminus (\text{Opt}_{n+1} \cup \text{Blo}_{n+1})$
  $n \leftarrow n + 1$
end while

Figure 4.1: The BIOT-algorithm

Blo_{n+1}, for all $n \geq 0$. It is moreover relatively easy to check that the BIOT-algorithm in figure 4.1 computes all and only weakly optimal pairs.

Proposition 4.1.1. If the BIOT-algorithm terminates in round n with $\text{Pool}_n = \emptyset$, then $\langle m, t\rangle \in \text{Opt}_n$ iff $\langle m, t\rangle$ is weakly optimal.

Proof. Let $g, g' \in \mathit{Gen}$ be arbitrary elements of the generator and n be the smallest number for which $\text{Pool}_n = \emptyset$. First, we will show that $g \in \text{Opt}_n$ implies that g is weakly optimal. Clearly, $\text{Opt}_k \subseteq \text{Opt}_{k+1}$ for any $1 \leq k \leq n$. So it suffices to show by induction that $g \in \text{Opt}_k$ implies g's weak optimality for $1 \leq k \leq n$. For $k = 1$, this is trivially so: if there are no better $g' \in \text{Gen}$ that differ from g only either along the form or the meaning dimension, then there are also no better weakly optimal g' with this property. Suppose therefore that all $g \in \text{Opt}_k$ are weakly optimal and suppose further, towards contradiction, that some newly added $g \in \text{Opt}_{k+1}$ that is not in Opt_k is not weakly optimal. If g is not weakly optimal, then there is some $g' \in \text{Gen}$ which shares with g either the form or the meaning component, which is weakly optimal, and is preferred to g. If g' is still in Pool_k, g is not in Opt_{k+1}. So either g' is blocked or optimal after round k. If g' is in Opt_k, g should no longer be in the pool, because it is blocked by g'. And if g' is in Blo_k, then there is some better form $g'' \in \text{Opt}_k$ (varying along either form or meaning dimension only), which by induction hypothesis is weakly optimal. But that means that g' cannot

be weakly optimal. Since this exhausts the space of possibilities, we have established that g is weakly optimal, which concludes the induction step.

It remains to the shown that Opt_n contains *all* weakly optimal pairs. Towards contradiction, assume that there is a weakly optimal g that is not in Opt_n. If $g \notin \text{Opt}_n$, then $g \in \text{Blo}_n$. But that means that there is some $g' \in \text{Opt}_n$ which varies from g only along either the meaning or form dimension such that $g' \succ g$. From the above we know that g' is weakly optimal. But if it is, g cannot be. □

4.1.3 Example: M-Implicatures in BiOT

Here is a simple example to illustrate how the BiOT-algorithm works. The example is BiOT's treatment of M-implicatures, as initially suggested by Blutner (Blutner 1998, 2000). We would like to explain why an unmarked form (9a) is paired with an unmarked meaning (9b), while a marked form (10a) is paired with a marked meaning (10b) (see also sections 1.1.2 and 2.2.2).

(9a) Black Bart killed the sheriff.

(9b) $\rightsquigarrow$ Black Bart killed the sheriff in a stereotypical way.

(10a) Black Bart caused the sheriff to die.

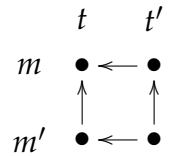
(10b) $\rightsquigarrow$ Black Bart killed the sheriff in a non-stereotypical way.

Let us assume that there are two forms m and m' corresponding with (9a) and (10a) respectively and two meanings t and t' representing (9b) and (10b). Initially, all four possible form-meaning pairs are in Gen. (This is because we assume that both forms are in principle compatible with either meaning.) We also assume that m is more costly than m' and that t is more stereotypical than t'. This gives rise to the following ordering $\succ$ over form-meaning pairs, basically two markedness constraints:

$$\begin{aligned} \langle m, \cdot \rangle &\succ \langle m', \cdot \rangle \\ \langle \cdot, t \rangle &\succ \langle \cdot, t' \rangle \end{aligned}$$

In words, m is preferred over m' independently of the associated meaning (because it is less costly to process), and t is preferred over t' independently of the associated form (because it is more stereotypical).

The initial situation, with which the algorithm starts, can be plotted as follows:



A bullet point • indicates that a form-meaning pair is in the pool, arrows between bullet points represent the ordering. Based on this initial configuration, the algorithm will compute the optimal pairs. In this case, $\langle m, t\rangle$ is the only optimal pair. This optimal pair will block —and therefore remove from the pool— the pairs $\langle m', t\rangle$ and $\langle m, t'\rangle$. The resulting situation after one round of iteration is:

$$\begin{aligned} \text{Opt}_1 &= \{\langle m, t\rangle\} \\ \text{Blo}_1 &= \{\langle m', t\rangle, \langle m, t'\rangle\} \\ \text{Pool}_1 &= \{\langle m', t'\rangle\} \end{aligned}$$

This can be represented as in the following diagram where a circled dot $\odot$ marks an optimal pair, and pairs no longer in the pool are crossed out $\times$:

	t	t'
m	$\odot$	$\times$
m'	$\times$	•

With only one pair left in the pool there is not much competition for optimality, so in the second round of iteration the BIOT-algorithm adds $\langle m', t'\rangle$ to the set of optimal pairs, removes it from the pool and terminates. The output of the algorithm is $\text{Opt}_2 = \{\langle m, t\rangle, \langle m', t'\rangle\}$ and the situation after two rounds of iteration looks like this:

	t	t'
m	$\odot$	$\times$
m'	$\times$	$\odot$

Under the above markedness constraints weak optimality thus predicts a unique mapping where m is paired with t and m' is paired with t'.[4]

4. With only the above markedness constraints, *strong* optimality does *not* predict that a marked form is associated with a marked meaning, since $\text{Opt}_1 = \text{BIOT}_{\text{str}} = \{\langle m, t\rangle\}$. Nevertheless, strong optimality can account for M-implicatures in full if we additionally assume a preference for associating marked forms with marked meanings, for instance in the form of additional so-called *harmony constraints*. However, this seems much less explanatory (see Blutner and Zeevat 2008, for discussion).

4.1.4 BiOT as a Model of Pragmatic Interpretation

The iterative BIOT-algorithm is certainly superficially reminiscent of the IBR model. BIOT's explanation of M-implicatures is also very much parallel to the treatment in the IBR model: a first iteration step deals with unmarked forms and meanings, and once this association is settled the actual M-implicature is accounted for, associating the marked form with the marked meaning. The main question to be explored in this chapter is therefore: how much of a parallel is there between BIOT on the one hand and IBR on the other? In order to address this question it is necessary to be clear about the conceptual interpretation of various optimality notions. What exactly does it mean when an OT-system selects a given form-meaning pair as weakly optimal but not strongly optimal, or as unidirectionally optimal but not strongly optimal?

Proponents of OT-pragmatics are not unanimous about this issue. Some propose to think of unidirectional and strong optimality as measures of online pragmatic competence, but reject the notion that weak optimality has anything to do with actual pragmatic reasoning (Blutner and Zeevat 2004, 2008). Weak optimality is rather viewed from a diachronic, evolutionary perspective as giving the direction into which the semantic meaning of expressions will most likely shift over time, by pragmatic pressures.

Opposed to this view, others treat also weak optimality as a model of pragmatic reasoning competence. Under this interpretation different notions of optimality express different levels of PERSPECTIVE TAKING: whereas unidirectional optimization does not require to take the interlocutor's perspective into account, bidirectional optimization does:[5]

> "[B]idirectional optimization requires the coordination of two opposite perspectives: the speaker's and the hearer's perspective. At the root of the mechanism of bidirectional optimization lies the assumption that the hearer takes into account which options the speaker has for expressing a given meaning, and that the hearer has some understanding of what makes the speaker choose a certain form. The latter assumption requires that the hearer takes into account that any choice the speaker makes is co-determined by the speaker's belief that the hearer will indeed be aware of these options. This means, first of all, that bidirectional optimization may require a child hearer to have a second-order theory of mind, and to be able to compute the implications of a recursive theory of mind."
>
> (Hendriks et al. 2007, section 5.6.2)

5. Hendriks et al. (2007) do not subscribe fully to this interpretation, but maintain it alongside other possible interpretations of optimality.

More strongly even, optimality theory in pragmatics is often related to theory of mind (ToM) reasoning (Premack and Woodruff 1978) (see also section 2.1.2). Unidirectional optimization is taken to involve no ToM reasoning (or zero-order ToM), strong optimization would correspond to first-order, and weak optimization would involve second-order ToM reasoning (see, for instance, Flobbe et al. 2008, p. 424).

Given the controversy about its conceptual interpretation, what would be required is, in a manner of speaking, an epistemic interpretation of optimality theory that clarifies (some of) its intended use in pragmatic applications. Thus conceived, a comparison to a related game theoretic model can help achieve this, especially when a game theoretic model has a proper epistemic interpretation, such as the IBR model does. This is what this chapter tries to achieve. I will eventually try to compare BiOT under the interpretation that different optimality notions express different competencies in perspective taking, to ToM reasoning in the vein of IBR.

Summary. In summary, a pragmatic OT-system abstractly defines preferences among possible form-meaning associations. There are then various notions of optimality which yield the predictions of the OT-system. This is reminiscent of the distinction between a game model on the one hand and various solution concepts on the other that was introduced in section 1.2. The questions that this comparison raises are (i) which game exactly a pragmatic OT-system corresponds to and (ii) which solution concept (together with a possible epistemic characterization) the different notions of optimality instantiate. The next section summarizes the 'received wisdom' on the matter.

4.2 BiOT and Game Theory

Bidirectional optimization is simultaneous optimization of both the production and the comprehension perspective. At first glance, this looks very similar to an equilibrium state in which the speaker's and the hearer's preferences are balanced. And, indeed, there is a *prima facie* very plausible link between BiOT and game theory. Dekker and van Rooij (2000) (henceforth D&vR) show that the notion of strong optimality corresponds one-to-one to the notion of Nash equilibrium in an *optimality game*.[6] An optimality game is a straight-

6. D&vR use the term "interpretation games" for what I call "optimality games." The former term would be equivocal in the context of this thesis, so I use the latter.

forward translation of an OT-system into a strategic game. D&vR continue to show that weak optimality corresponds to the outcome of a process that we could call *iterated Nash-selection*. Let's first look at the analysis of D&vR in more detail and then reflect critically.

4.2.1 BiOT and Strategic Games

Optimality Games. Recall from section 1.2.1 that a strategic game is a triple $\langle N, (A)_{i\in N}, (\succeq)_{i\in N}\rangle$ where N is a set of players, A_i are the actions available to player i and $\succeq_i$ is player i's preference relation over action profiles $\times_{j\in N} A_j$, i.e., possible outcomes of the game. A Nash equilibrium of a strategic game is an action profile a^* such that for all $i \in N$ there is no $a_i \in A_i$ for which:

$$(a^*_{-i}, a_i) \succ_i a^*.$$

Take an OT-system with forms M, meanings T —assuming for simplicity that Gen $= M \times T$— and some ordering $\succeq$ over form-meaning pairs. An optimality game, as defined by D&vR, is a strategic game between a speaker S and a hearer H such that the speaker selects a form, $A_S = M$, the hearer selects a meaning, $A_H = T$, and the players' preferences are just equated with the ordering of the OT-system, $\succeq_S = \succeq_H = \succeq$.

Strong Optimality as Nash Equilibrium. An action profile $\langle m, t\rangle$ is a Nash equilibrium of an optimality game iff

(i) there is no $m' \in M$ such that $\langle m', t\rangle \succ_S \langle m, t\rangle$; and

(ii) there is no $t' \in T$ such that $\langle m, t'\rangle \succ_H \langle m, t\rangle$.

But since $\succeq_S = \succeq_H = \succeq$ this is the case just when $\langle m, t\rangle \in \text{BIOT}_{\text{str}}$. Consequently, every Nash equilibrium of an optimality game is a strongly optimal pair in the corresponding OT-system, and every strongly optimal pair of an OT-system is a Nash equilibrium of the corresponding optimality game. D&vR's result in slogan form: strong optimality is Nash equilibrium (in an optimality game).

Weak Optimality as Iterated Nash-Selection. D&vR's characterization of weak optimality is inspired by the BiOT-algorithm given in section 4.1.2. Recall that the BiOT-algorithm iteratively computes strongly optimal pairs, based on a shrinking pool of candidate pairs. Since strong optimality can be likened to Nash equilibrium in optimality games, the workings of the BiOT-algorithm can be recast in game theoretic terms as a process of iteratively

removing action profiles from competition for Nash equilibrium that are, in a way of speaking, dominated by a Nash equilibrium.

In order to make this idea more precise, D&vR allow strategic games to have partial preferences. For games with partial preferences, not every definition of Nash equilibrium will do, but the one given above applies. The process of ITERATED NASH-SELECTION on a strategic game $I_0 = \langle N, (A)_{i \in N}, (\succeq_0)_{i \in N} \rangle$ is defined inductively as follows: let NE_n be the set of Nash equilibria of game I_n; I_{n+1} is derived from I_n by restricting the preferences $\succeq_{n,i}$ to:

$$\succeq_{n+1,i} = \{ \langle x, y \rangle \in \succeq_{n,i} \mid \neg \exists z \in \text{NE}_n : z \succ_{n,i} x \} .$$

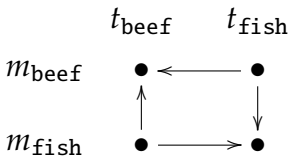
If for some index n we have $I_n = I_{n+1}$, we consider the process to be terminated, and call NE_n the *outcome* of the process of iterated Nash-selection. D&vR show that this process corresponds to the BIOT-algorithm if applied to optimality games: if I is the optimality game corresponding to an OT-system, then the outcome of iterated Nash-selection on I contains all and only the weakly optimal pairs of the OT-system.

4.2.2 Critique

The characterization of strongly optimal pairs as Nash equilibria in an optimality game has some *prima facie* plausibility and seems unanimously endorsed as *the* link between OT and game theory by the pragmatic OT community. But on closer look the suggested parallel turns out not to be very sensible. Moreover, although weak optimality has a very tight correspondence via the BIOT-algorithm with the process of iterated Nash-selection, this latter is not a standard solution procedure in game theory — and that is so for a good reason. This is what the following lines will argue for, with the conclusion that the true connection between optimality theoretic and game theoretic pragmatics is still an open issue.

ON OPTIMALITY GAMES. I would like to argue first that the translation of a pragmatic OT-system into a strategic optimality game is dubious. This point is best made based on a simple example. Here is a very simple OT-system that captures the wine-choice scenario from section 1.2.2 where Alice would like to inform Bob whether it's beef or fish for dinner. There are two forms $m_{\texttt{beef}}$ and $m_{\texttt{fish}}$ and two meanings $t_{\texttt{beef}}$ and $t_{\texttt{fish}}$. As for the ordering, let's only require that form-meaning pairs are subject to Gricean Quality as a faithfulness constraint: any pair $\langle m, t \rangle$ where m is semantically compatible with t

is strictly preferred over any pair where this is not so. We thus obtain the following OT-system for the obvious meanings of $m_{\texttt{beef}}$ and $m_{\texttt{fish}}$:[7]



The strongly optimal pairs are $\langle m_{\texttt{beef}}, t_{\texttt{beef}} \rangle$ and $\langle m_{\texttt{fish}}, t_{\texttt{fish}} \rangle$, and this clearly is the only reasonable prediction with hardly a reason to worry. But let's now have a critical look at what it means to imagine that Alice and Bob are playing a *strategic* optimality game here. It would mean that Alice has to make a decision what to say *irrespective* of the actual state, while Bob chooses a meaning *independently* of Alice's choice. It is quite clear that this is not the correct analysis of an informative utterance and its possible interpretation (compare the argument in section 1.2.2): speakers do not choose a form irrespective of the idea that they want to express, and hearers do not choose an interpretation irrespective of a form that they want to interpret. But this is exactly what it means to play an optimality game. The characterization of an OT-system as a strategic game is not faithful to our intuitions about the temporal and informational dynamics of an utterance and its uptake. Therefore, if an OT-system is to serve as a model of pragmatic interpretation, an analysis in terms of a strategic game seems dubious.

On Nash Equilibrium. This has repercussions for the analysis of strongly optimal pairs as Nash equilibria. Under the standard interpretation of Nash equilibrium as a *steady state* in the behavior of agents when playing a game recurrently (see section 1.2.1), we predict Alice and Bob to have settled by force of precedent on, for instance, saying $m_{\texttt{beef}}$, no matter what is actually prepared for dinner, and always taking it that beef is going to be prepared, irrespective of any observed message. Clearly, not only the interpretation of the communicative situation as a strategic game, but also the interpretation of a strongly optimal pair as a Nash equilibrium is inadequate.

On Iterated Nash-Selection. What should we then say of the characterization of weak optimality in terms of iterated Nash-selection? Obviously, the

7. It does not matter for the argument at hand that there are no attested pragmatic inferences in this scenario. We could also assume that only pairs $\langle m, t \rangle$ are allowed in the generator such that m is true in t. The gist of the argument remains. In fact, my point could equally well be made based on *any* arbitrary OT-system that has more than one strongly optimal pair.

latter is the direct translation of the BIOT-algorithm into game theory, if we assume that a pragmatic OT-system should be analyzed as a strategic game. Despite the above arguments why this is not a good analysis, we can still review arguments for or against iterated Nash-selection independently. This procedure is not standard in game theory, and so the question arises why.

The simple answer is: iterated Nash-selection is not an attractive solution procedure because it crucially hinges on but strictly goes against the idea of a Nash equilibrium as the solution concept of a strategic game. In classical game theory, Nash equilibrium is used as a predictor of how instrumentally rational agents would or ought to behave in a repeated situation of strategic interaction. Behavioral explanations in terms of Nash equilibrium then face the difficulty that the presence of multiple equilibria undermines a unique prediction. This dilemma also arose for GTP, as we have seen in sections 1.2.3 and 1.2.4: we wanted to restrict the set of possible equilibrium solutions, not to make it even larger. So conceived, iterated Nash-selection only makes matters worse: it produces *even more* equilibria, even in a case, like the M-implicature example in section 4.1.3, where there is initially exactly one, beautifully unique prediction of Nash equilibrium. That is why classical game theory would not endorse iterated Nash-selection.

4.2.3 BIOT and Signaling Games

The above considerations suggest that the natural way of interpreting a set of form-meaning pairs —be they optimal or not— is not as a set of Nash equilibria, but rather as a (possibly partial) specification of a sender or a receiver strategy in a signaling game. Consider again the set of optimal pairs in the wine-choice example of the last section. In this simple example, all four notions of optimality coincide and yield the same two form-meaning pairs as the prediction of the OT-system:

$$\{\langle m_{\texttt{beef}}, t_{\texttt{beef}}\rangle, \langle m_{\texttt{fish}}, t_{\texttt{fish}}\rangle\}.$$

How should we interpret this prediction? Obviously, this set specifies the speaker's optimal production behavior and the receiver's optimal comprehension behavior: it specifies that the speaker would optimally choose $m_{\texttt{beef}}$ whenever she wants to express the meaning $t_{\texttt{beef}}$ and $m_{\texttt{fish}}$ when she wants to express the meaning $t_{\texttt{fish}}$, and that the hearer would optimally interpret $m_{\texttt{beef}}$ as meaning $t_{\texttt{beef}}$ and $m_{\texttt{fish}}$ as meaning $t_{\texttt{fish}}$. But this means that form-meaning pairs should not be looked at individually but rather interpreted *as*

a set that specifies a *function*: a set of optimal form-meaning pairs should be linked to a *strategy*, i.e., a specification of *conditional* behavior, in a suitable dynamic game.

In particular, a set of form-meaning pairs partially defines a sender or receiver strategy in a SIGNALING GAME WITH INTERPRETATION ACTIONS where

(i) the set of states in the signaling game are the meanings T of the OT-system; these are the meanings that the speaker might want to express;

(ii) the set of messages in the signaling game are the forms M of the OT-system; these are the messages the speaker can choose to express a meaning when she wants to; and

(iii) the set of receiver actions in the signaling game are interpretations, i.e., the meanings T of the OT-system.

In general, we can read off a (partial) description of a sender and receiver strategy for such a game from any set $O \subseteq M \times T$. The set of pure sender strategies in a signaling game with interpretation actions compatible with O is:[8]

$$\mathsf{S}(O) = \{s \in \mathsf{S} \mid O(t) \neq \emptyset \rightarrow s(t) \in O(t)\}\,;$$

and the set of pure receiver strategies compatible with O is:

$$\mathsf{R}(O) = \{r \in \mathsf{R} \mid O(m) \neq \emptyset \rightarrow r(m) \in O(m)\}\,.$$

Obviously, an arbitrary set O need not specify a full strategy. There may be states t for which $O(t)$ is empty, so that when taken as a description of a sender strategy O is only a *partial* description. I suggest that this is really how we should set the link between OT and game theory in pragmatics: sets of form-meaning pairs —no matter whether any notion of optimality has selected these— are specifications of strategies in a corresponding signaling game with interpretation actions.

8. Recall that we use the following notation:

$$\begin{aligned} O(t) &= \{m \in M \mid \langle m, t\rangle \in O\} \\ O(m) &= \{t \in T \mid \langle m, t\rangle \in O\}\,. \end{aligned}$$

4.3 An Epistemic Interpretation of Optimality

Natural as it may be, linking form-meaning pairs to strategies does not yet fix a complete translation between OT-systems and signaling games. Some correspondences are hardly worth mentioning. Speakers correspond to senders and hearers correspond to receivers, of course. The generator places restrictions on the set of possible form-meaning associations and this naturally finds its expression in the semantic denotation function

$$\langle m, t \rangle \in \text{Gen iff } t \in [\![m]\!]$$

if we assume that the corresponding signaling game makes truthful signaling obligatory. This leaves us with the ordering $\succeq$ of the OT-system, and three elements of the signaling game left to be matched and/or somehow specified: the prior probabilities $\Pr(\cdot)$, and the utilities $U_{S,R}$ for both sender and receiver.

Formally, there are many possibilities of translation between OT-systems and signaling games. Which formal possibility is most sensible depends on the intended application of BiOT. Recall from section 4.1.4 that the prevalent interpretation of BiOT, if considered a description of online pragmatic competence, is that unidirectional optimization involves no perspective taking, but that bidirectional optimization does. In this section I would like to address this interpretation of optimality critically, with a comparison of OT and IBR. To motivate my formal comparison, I will first discuss a case study in section 4.3.1 showing how BiOT is applied to data from language acquisition, in particular comprehension/production mismatches in early acquisition. This is to set the scene, motivate and exemplify the way a notion of "perspective taking" is employed in BiOT for explanatory purposes. I will then argue that optimality notions should be linked to strategic types of, in particular, the R_0-sequence of the IBR model. This yields an interesting epistemic characterization of optimality notions as follows: unidirectional optimality is Bayesian rationality in its most basic form; strong optimality corresponds to one round of perspective taking of a naïvely updating receiver; and weak optimality is the limit behavior of a receiver who adheres to the BiOT-algorithm's conservative notion of blocking and optimality.

4.3.1 Comprehension Lags in Language Acquisition

When a young child learns its first language, common sense might expect that competence in comprehension temporally precedes competence in production (cf. Smolensky 1996): after all, how should a language-learning child

be able to use correct expressions in the right circumstances, when it isn't even able to understand these forms properly when it hears them? In general, any mismatch in comprehension or production competences during acquisition challenges a theory of grammar, because it needs to be explained how it is possible to use grammatical competence correctly in one way, but not in another way. When production lags behind comprehension, an explanation in terms of insufficient computational resources, such as working memory or planning capacity, might seem (relatively) ready at hand. But COMPREHENSION LAGS, i.e., examples where children first acquire competence in production and only later in comprehension, are not quite as easy to explain in terms of computational demands: it is much more plausible to assume that 'active' production is a more resource-intensive process than mere 'passive' comprehension, or so it would seem.

Nonetheless, there are numerous examples of comprehension lags, such as in the interpretation of reflexive and non-reflexive pronouns (Hendriks and Spenader 2005), the interpretation of indefinites (de Hoop and Krämer 2005), or the interpretation of contrastive stress (Hendriks et al. 2007).[9] It pays to zoom in on only one of these examples in some detail, so as to understand the general pattern of explanation and to pick up the main idea for subsequent discussion and comparison to a game theoretic approach.

THE PRONOUN INTERPRETATION PROBLEM. Hendriks and Spenader (2005) discuss the following astonishing comprehension lag concerning the meaning of reflexive pronouns. Clearly, for (most) adult speakers of English the sentence (59) has only a coreferential reading for the reflexive pronoun, i.e., (59) means that Bert washed Bert (and not Ernie or any nearby male yellow rubber duck). In contrast, sentence (60) has *no* coreferential reading for the non-reflexive pronoun, i.e., (60) means that Bert washed someone other than himself.[10]

(59) Bert washed himself.

(60) Bert washed him.

Young children, on the other hand, have difficulties with these sentences and show a peculiar pattern of production and comprehension asymmetry in early

9. For general discussion see also Hendriks et al. (2007) and Hendriks (2008).

10. I would like to encourage the reader not to get carried away too far in imagining possible referents of the pronoun in (60). In laboratory experiments, there would just be two salient referents: Ernie and Bert.

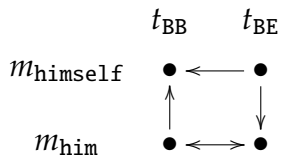
acquisition (see Hendriks and Spenader 2005, for details and further references). In comprehension, up to 95% of 3-year-olds assign a correct coreferential reading to (59), but about half the children of this age group wrongly assign to (60) a coreferential reading as well. By the age of 6-7, however, comprehension of these sentences matches adult competence. In contrast, production equals adult competence already at the earlier stage of language acquisition. This data poses the interesting question how it is possible that young children's grammatical knowledge and their general computational abilities enable (i) adult-like production of both forms (59) and (60), (ii) adult-like comprehension of (59), but (iii) improper comprehension of (60)?

A BiOT Account of Pronoun Interpretation Data. Hendriks and Spenader propose that this asymmetry originates in the inability of young interpreters to reason about alternative forms the speaker could have used. This idea is spelled out in a model of grammatical competence in the form of an OT-system with two forms $m_{\texttt{himself}}$ for (59) and $m_{\texttt{him}}$ for (60), and two meanings $t_{\texttt{BB}}$ for a situation in which Bert washed Bert and $t_{\texttt{BE}}$ for a situation in which Bert washed Ernie. All possible form-meaning combinations are generated in this system and the ordering is derived from two constraints:

Principle A: a faithfulness constraint that gives preference to coreferential readings of reflexives: only the pair $\langle m_{\texttt{himself}}, t_{\texttt{BE}} \rangle$ violates this constraint; and

Referential Economy: a markedness constraint on forms that prefers reflexive pronouns over non-reflexive pronouns: both pairs $\langle m_{\texttt{him}}, \cdot \rangle$ violate this constraint once.

It is assumed that Principle A outranks Referential Economy, which results in an OT-system that can be vizualized as follows:



The ordering gives rise to the following sets of optimal pairs (the notation is suggestive of my preferred functional interpretation of these sets):

$$\text{Opt}_{\text{syn}} = \left\{ \begin{array}{lcl} t_{\texttt{BB}} & \mapsto & m_{\texttt{himself}} \\ t_{\texttt{BE}} & \mapsto & m_{\texttt{him}} \end{array} \right\}$$

$$\text{Opt}_{\text{sem}} = \left\{ \begin{array}{lcl} m_{\texttt{himself}} & \mapsto & t_{\texttt{BB}} \\ m_{\texttt{him}} & \mapsto & t_{\texttt{BB}}, t_{\texttt{BE}} \end{array} \right\}$$

$$\textsc{biot}_{\text{str,weak}} = \left\{ \begin{array}{lcl} t_{\texttt{BB}} & \leftrightarrow & m_{\texttt{himself}} \\ t_{\texttt{BE}} & \leftrightarrow & m_{\texttt{him}} \end{array} \right\}$$

This fits the acquisition data beautifully: young children's comprehension and production behavior may be mapped onto unidirectional optimization, while adult-like performance corresponds to bidirectional optimization. Hendriks and Spenader propose that this models the young child's inability to take the speaker's perspective, in particular her expression alternatives, into account.

4.3.2 Unidirectional Optimality

Suppose we accept Hendriks and Spenader's explanation of the comprehension lag in pronoun interpretation. Suppose also that we accept my characterization of sets of form-meaning pairs as partial strategies in signaling games with interpretation actions. The relevant question then is: how do we translate unidirectional and bidirectional optimality into a game theoretic model in a way that respects the spirit of Hendriks and Spenader's explanation, i.e., in a way that respects the idea that young children fail bidirectional interpretation because they do not take the speaker's options into consideration? The obvious idea is to assume that young language learners have not yet acquired either the skills or the resources to perform higher-level reasoning in the IBR model. Unidirectionally optimal behavior maps onto lower-level strategic types. Bidirectionally optimal behavior maps onto higher-level strategic types. The question then becomes: which types exactly?

Unidirectional Optimality as Behavioral Bias. A possibility that I would like to raise, only to dismiss it eventually, is to match unidirectional optimality with the behavior of level-zero players in the obvious sense that $\text{Opt}_{\text{syn}} = S_0$ and $\text{Opt}_{\text{sem}} = R_0$. The problem I have with this idea is that it requires amendment of the IBR model. To see this, consider the above OT-system for the pronoun interpretation puzzle. As it stands, since S_0 is defined to send arbitrary true messages in each state, the only way of matching S_0's behavior

with Opt_{syn} is to assume that $[\![m_{\texttt{himself}}]\!] = \{t_{\texttt{BB}}\}$ and that $[\![m_{\texttt{him}}]\!] = \{t_{\texttt{BE}}\}$. This is clearly *not* a plausible assumption for a signaling game model of this scenario, but even if we made this contestable assumption, it would not be possible to match R_0's behavior to Opt_{sem}, because if $[\![m_{\texttt{him}}]\!] = \{t_{\texttt{BE}}\}$ then it is not possible that $t_{\texttt{BB}} \in R_0(m_{\texttt{him}})$ under the given definition of the IBR types.

To maintain an interpretation of unidirectional optimality with level-zero players, we would therefore have to redefine the beginning of the IBR sequences. For instance, we could assume that S_0 is unstrategically sending only true messages but is moreover susceptible to message costs. This way it is possible to match Opt_{syn} for the pronoun interpretation puzzle by assuming, as is usual in signaling games, that costs of messages depend on states. Slightly differently, but to a similar effect, we might also assume that the constraints specified by a given OT-system are additional *grammatical biases* of otherwise unstrategic level-zero players: thus conceived, we would implement not only truth-conditional meaning but more complex syntactico-semantic features such as non-binding focal elements into a pragmatic reasoning system.

I find especially this latter idea plausible enough and even appealing and promising: after all, considering further grammatical biases as focal reasoning points might open up the game theoretic model to new realms of application. Still, I will not pursue this approach any further here, because, firstly, for the time being I would prefer a more conservative comparison of BIOT with the IBR model as it stands, and, secondly, there is another plausible alternative approach to comparison for which we do not have to change the basic definitions of the IBR model.

Unidirectional Optimality as Least Sophisticated Optimization. Although unidirectional optimization does not take the other interlocutor's perspective into account, it is nonetheless a process of optimization. This suggests that we should match unidirectional optimality with the least sophisticated strategic types in the IBR model that do perform some kind of optimization. These are, interestingly enough, R_0 and S_1.[11] Matching $\text{Opt}_{\text{sem}} = R_0$ and $\text{Opt}_{\text{syn}} = S_1$ implies that the ordering of a given OT-system gives the expected utilities of R_0 and S_1 respectively. Thus conceived, the question is whether

11. This is an interesting point to notice in passing: level-zero senders need not behave rationally at all; level-zero receivers, on the other hand, behave rationally given a possibly irrational belief in literal interpretation. Whence the asymmetry in players' optimization behavior.

	$\Pr(t)$	$a_{\texttt{BB}}$	$a_{\texttt{BE}}$	$m_{\texttt{himself}}$	$m_{\texttt{him}}$
$t_{\texttt{BB}}$	$1/2$	1,1	0,0	$\surd$	$\surd$
$t_{\texttt{BE}}$	$1/2$	0,0	1,1	–	$\surd$

Figure 4.2: Signaling game for the pronoun interpretation puzzle

this always necessarily yields a full translation of an OT-system into a signaling game.

A straightforward translation is possible for the pronoun interpretation puzzle. This behavior of agents falls out under a standard definition of the IBR model for a signaling game like in figure 4.2. However, it may be objected here that to assume $[\![m_{\texttt{himself}}]\!] = \{t_{\texttt{BB}}\}$ is unwarranted and not what the corresponding OT-system would do. If this is perceived as a problem, an alternative way of setting up the signaling game is conceivable. If we assume that $[\![m_{\texttt{himself}}]\!] = \{t_{\texttt{BB}}, t_{\texttt{BE}}\}$, then we need to make sure that R_0 still matches Opt_{syn}. This is possible if we take recourse to the idea that prior probabilities in an interpretation game are only a compact way of specifying posterior probabilities (section 3.1). Whenever this compact representation proves too restricted, as in the present case, we may wish to resort to a different, more flexible specification. It is thus compatible with the interpretation of the context model and the standard IBR model to assume that $\Pr(t_{\texttt{BB}}|m_{\texttt{himself}})$ is bigger than $\Pr(t_{\texttt{BE}}|m_{\texttt{himself}})$ while $\Pr(t_{\texttt{BB}}|m_{\texttt{him}})$ is equal to $\Pr(t_{\texttt{BE}}|m_{\texttt{him}})$. In fact, if we allow for this latter more flexible specification of posterior beliefs of the receiver it is immediate that there is always a signaling game model that corresponds to any given OT-system in the sense that the sets of unidirectionally optimal form-meaning pairs match R_0 and S_1. Such corresponding signaling games may have to assume sender response utilities and message costs quite uncharacteristic of interpretation games, and the translation from OT-system to signaling game model is not unique but one-to-many. Nonetheless, existence of a suitable signaling game is guaranteed.

Primacy of Production. The obvious criticism is that my suggested parallelism renders the sender strangely more sophisticated than the receiver. This, however, need not be implausible for a model of language use and interpretation. Some people see production as a more active, deliberate process than passive, reactive interpretation: for instance, Zeevat (2000) argues that there seems to be a natural primacy of production over comprehension in the sense

that even the most naïve form of intentional speaking is more of an *active* decision making than the most naïve form of listening. Zeevat therefore argues for an asymmetric approach to optimality notions and suggests a system that takes OT-syntax as its central axis around which cooperative pragmatic reasoning optimizes for both speaker and hearer. Thus conceived, by mapping the OT-ordering $\preceq$ to R_0 and S_1, Zeevat's asymmetric picture is compatible even with standard BIOT. Moreover the parallel between BIOT and IBR that I suggest here may perhaps even be taken as a *formal* plausibility argument for the priority of production over comprehension in a model of language competence: in the IBR model the least sophisticated speakers that optimize at all are more sophisticated than the least sophisticated optimizing hearers.

QUANTITY AS SEMANTIC STRENGTH. Further support for my translation proposal can be found in other applications of BIOT to pragmatics. In early work, Blutner (1998) applied OT —though it was not yet identified as such at the time— to the computation of conversational implicatures. Towards this end, Blutner assumed that each form $m \in M$ was associated with a cost $c(m) > 0$, and a meaning $[\![m]\!] \subseteq T$. Blutner then defined the ordering $\succeq$ directly in terms of a function $C : \text{Gen} \to \mathbb{R}$ as:

$$g \succeq g' \text{ iff } C(g) \geq C(g')$$

where

$$C(m,t) = c(m) \times -\log_2 \Pr(t|\,[\![m]\!]).$$

Blutner left the prior probabilities of states unanalyzed, but the main idea behind his approach is apparent: a form-meaning pair is relatively preferred the cheaper the form is, and the more likely the meaning is given that the form is true. For a fixed form, the hearer ordering $\succeq_m$ will select the most likely meaning given the semantic meaning of the message. For a fixed meaning, the speaker ordering $\succeq_t$ will select the form which at the same time minimizes the costs and maximizes the likelihood of the to-be-expressed meaning.

It is easy to see that Blutner's hearer ordering implements the expected utility of a level-zero receiver in an interpretation game:

$$\begin{aligned} t \succeq_m t' \quad &\text{iff} \quad \Pr(t|\,[\![m]\!]) \geq \Pr(t'|\,[\![m]\!]) \\ &\text{iff} \quad \mathrm{EU}_R(t,m,\mu_0) \geq \mathrm{EU}_R(t',m,\mu_0). \end{aligned}$$

Similarly, Blutner's speaker ordering implements the expected utility of a level-1 sender in an interpretation game with flat priors if we assume that

costs $c(m)$ are infinitesimally small, so as to assimilate nominal costs. In general, Blutner's ordering contains the idea that the speaker prefers a true form m over another true and equally costly form m' if (but not necessarily only if) $[\![m]\!] \subset [\![m']\!]$. But that means that Blutner's ordering implicitly implements a speaker conjecture that the receiver is interpreting literally: only under an expectation of literal uptake devoid of pragmatic inference is it always rational to prefer semantically stronger statements all else being equal.

A similar point can be made in connection with the constraint-based pragmatic OT-system initiated by Aloni (2007), which is further developed by Pauw (2008). Both Aloni and Pauw translate Gricean Maxims rather directly into constraints of a pragmatic OT-system. Quantity is implemented as a constraint that strictly prefers a form m over a form m', all else being equal, just in case $[\![m]\!] \subset [\![m']\!]$. Just as before, we again discern here the hidden assumption —in essence a speaker conjecture— that the hearer is naïve and interprets forms based on their semantic meaning only.

The insight that emerges here is actually noteworthy in general. It's interesting to ask why we would like to implement Gricean Quantity in terms of semantic strength. Why is it that a more *informative* message is one that is *semantically* more specific? After all, one could imagine that the requirement to be informative aims at the *outcome* of communication rather than the input. To wit: if you manage to understand me perfectly, even if I use a tautology —semantically totally uninformative— and even better than when I had used any semantically stronger sentence, why should I still prefer semantically stronger messages? My argument therefore is that whenever a pragmatic approach implements Gricean Quantity, as relevant BIOT approaches have done, in terms of semantic strength (as opposed to in terms of any further systematic pragmatic enrichment), these approaches are implicitly assuming that speakers are (something like) level-1 players that rely on literal, non-enriched uptake in their optimization. Whence that it is legitimate to link Opt_{syn} with S_1 and Opt_{sem} with R_1.

4.3.3 Bidirectional Optimality

If unidirectional optimality corresponds to the behavior of R_0 and S_1, what then does bidirectional optimality correspond to? It is certainly not farfetched to suspect that strong optimality might coincide with R_2's interpretation behavior, and that the iterating BIOT-algorithm, which computes weak optimality, just corresponds to the interpretation behavior of higher-level re-

ceiver types in the R_0-sequence. This suggestive idea is made even more plausible if we compare the way generalized M-implicatures are computed in both systems. If after i rounds of computation of the BIOT-algorithm a form m_j, $j \leq i$, is in the set of optimal form-meaning pairs, then we find $R_{2i}(m_j) = \{t_j\} = \text{Opt}_i(m_j)$, at least if we assume divine k-dominance. With only weak k-dominance or no FI assumption at all we still get that if after i rounds of computation of the BIOT-algorithm a form m_j, $j < i$, is optimal, then $R_{2i}(m_j) = \{t_j\} = \text{Opt}_i(m_j)$ and $R_{2i}(m_i) \supseteq \{t_i\} = \text{Opt}_i(m_j)$.

The following results show that this plausible conjecture is only almost correct: if we assume that the OT-system can be translated so that its ordering corresponds to the expected utility of R_0 and S_1 in an interpretation game then we can show that strong optimality is characterized by higher-order receiver types that perform certain, restricted kinds of belief update. In particular, strong optimality is equivalent to the interpretation of an *unsophisticated* level-2 receiver as introduced in section 2.2.3. Weak optimality, on the other hand, is not equivalent to the limit behavior of unsophisticated receivers. To match the behavior of the BIOT-algorithm in IBR we need to assume an even more restricted form of receiver belief formation. Weak optimality is equivalent to the interpretation behavior of a *myopic* receiver who strongly adheres to a strictly conservative notion of optimality: once a form is associated with a given meaning, a myopic receiver will always adhere to this association in forming his posterior beliefs. Together, these results then give an epistemic characterization of bidirectional optimality within the IBR model in terms of different kinds of belief formation of the receiver.

Strong Optimality as Unsophisticated Update

Section 2.2.3 elaborated on the difference between sophisticated and naïve posterior belief formation of the receiver. Recall that a receiver of level k updates naïvely if he adopts posteriors of the form

$$\mu_k(t|m) = \Pr(t|S_{k-1}(m)),$$

where

$$S_{k-1}(m) = \{t \in T \mid \exists s \in S_{k-1} : \; s(t) = m\}.$$

An unsophisticated receiver R_k assumes —possibly inconsistently— that all types $t \in S_{k-1}(m)$ that may send a given message m according to the behavioral belief S_{k-1} *always* send this message.

If we equate the ordering of an OT-system with the expected utility of R_0 and S_1 in an interpretation game, then strong optimality can be characterized as the interpretation behavior of an unsophisticated level-1 receiver.

Proposition 4.3.1. If for some message $\textsc{biot}_{\text{str}}(m) \neq \emptyset$, then $\textsc{biot}_{\text{str}}(m) = R_2(m)$ if R_2 performs an unsophisticated update.

Proof. Let $t \in \textsc{biot}_{\text{str}}(m)$. Then m is a rational choice for S_1 in state t. That implies that m is not a surprise for R_2, so that $\mu_2(\cdot|m)$ is derived from naïve consistency as $\mu_2(\cdot|m) = \Pr(t|S_1(m))$. Next, since $t \in \textsc{biot}_{\text{str}}(m)$, we also know that $t \in R_0(m)$, which means that t maximizes $\Pr(\cdot|\,[\![m]\!])$. But then t also maximizes $\Pr(\cdot|S_1(m))$, given that S_1 will not send untrue messages by assumption.[12] But that just means that $t \in R_2(m)$.

As for the other direction of inclusion, assume that $\textsc{biot}_{\text{str}}(m) \neq \emptyset$ and $t \in R_2(m)$. The former implies that $S_1(m) \neq \emptyset$ and the latter just means that t maximizes $\Pr(\cdot|S_1(m))$. Together this yields that $t \in S_1(m)$, or alternatively that it's rational for S_1 to send m in t. It then remains to show that it's also rational for R_0 to interpret m as t, i.e., we need to show that t maximizes $\Pr(\cdot|\,[\![m]\!])$. Towards this end, observe that there is some $\bar{t} \in \textsc{biot}_{\text{str}}(m)$ by assumption, which is therefore maximal in $\Pr(\cdot|\,[\![m]\!])$. From the above we know that $\bar{t} \in R_2(m)$. But that means that $\Pr(t|S_1(m)) = \Pr(\bar{t}|S_1(m))$. This implies that $\Pr(t) = \Pr(\bar{t})$ which in turn implies that t also maximizes $\Pr(\cdot|\,[\![m]\!])$ if we assume that S_1 sends only true messages. Together we obtain that $t \in \textsc{biot}_{\text{str}}$. □

Why Naïvety is Necessary. For a characterization of strong optimality as level-2 interpretation behavior, the restriction to unsophisticated update behavior of the receiver is necessary. To see where sophisticated update differs from strong optimality, consider the some-all game with skewed priors that was discussed already in section 2.2.4. The example was a some-all signaling game with $\frac{2}{3} > \Pr(t_\forall) > \frac{1}{2}$. With these prior probabilities the R_0-sequence starts as follows:

$$R_0 = \left\{ \begin{array}{ccc} m_{\texttt{some}} & \mapsto & t_\forall \\ m_{\texttt{all}} & \mapsto & t_\forall \end{array} \right\} \qquad S_1 = \left\{ \begin{array}{ccc} t_{\exists\neg\forall} & \mapsto & m_{\texttt{some}} \\ t_\forall & \mapsto & M \end{array} \right\}$$

A sophisticated R_2 would respond to $m_{\texttt{some}}$ under the belief in S_1 by properly taking into account that S_1 also sends $m_{\texttt{all}}$ in state $t_\forall$. With $\frac{2}{3} > \Pr(t_\forall) > \frac{1}{2}$

12. Remember that this is how we implemented the restrictions of the generator.

this yields:

$$R_2 = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_{\exists\neg\forall} \\ m_{\texttt{all}} & \mapsto & t_{\forall} \end{array} \right\}.$$

Opposed to that, an unsophisticated R_2 plays:

$$R_2' = \left\{ \begin{array}{lcl} m_{\texttt{some}} & \mapsto & t_{\forall} \\ m_{\texttt{all}} & \mapsto & t_{\forall} \end{array} \right\}.$$

This is because a naïve R_2 updates his priors with the set of states in which S_1 would ever send a given message, which in the case of $m_{\texttt{some}}$ is the set of all states.

Strong optimality does not model the sophisticated update behavior, but follows the unsophisticated receiver. The set of strongly optimal pairs in this example are:

$$\textsc{biot}_{\text{str}} = \left\{ \langle m_{\texttt{some}}, t_{\forall} \rangle , \langle m_{\texttt{all}}, t_{\forall} \rangle \right\}.$$

Roughly speaking, since strong optimality merely computes the intersection of strategies of R_0 and S_1, it is not sensitive to the kind of distributional information —which message is sent in how many states— that a sophisticated updater takes into account.

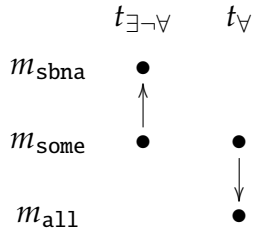
Weak Optimality as Myopic Update

Let's briefly recapitulate some of the previous results. Strong optimality specifies the behavior of unsophisticated level-2 receivers. Furthermore, strong optimality is also the set Opt_1 of optimal form-meaning pairs after one round of iteration of the BIOT-algorithm. Since moreover the BIOT-algorithm is a repeated application of strong optimality after removal of optimal and blocked form-meaning pairs, it is tempting to suspect that the set of optimal interpretations Opt_i after i rounds of iterations partially characterizes the behavior of level-2i receivers, if we assume that the receiver performs an unsophisticated update throughout.

This idea is not correct. It turns out that the BIOT-algorithm actually is peculiarly conservative: the monotonicity of the sets Blo_n and Opt_n means that (i) if a form-meaning pair is blocked, it will be completely removed from further consideration, and that (ii) if $\langle m, t \rangle$ is selected as optimal at some round, the association between m and t is fixed for good and always. The IBR model, on the other hand, whether defined with sophisticated or unsophisticated receivers, does *not* generally block or fix form-meaning associations once and

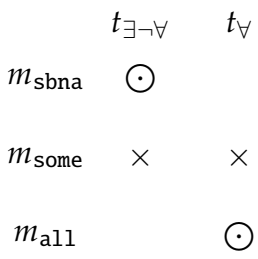
for all. Here are two examples that show, respectively, how (i) the IBR model can select interpretations in later rounds that the BIOT-algorithm has discarded as blocked, and how (ii) the IBR model can rule out an interpretation that the BIOT-algorithm has selected as optimal in earlier iterations.

THE SYMMETRY PROBLEM AGAIN. Consider anew the simple extension of the some-all game that we have looked at before, in section 2.3, where we assume that an additional third form $m_{\texttt{sbna}}$, short for "some but not all," is given which has the obvious semantics but which also incurs a slight cost. Translating this constellation into an OT-system would yield the following initial constellation under the Blutner ordering:[13]



This is also exactly what a direct translation from R_0's and S_1's expected utilities would yield.

The optimal pairs after the first round of iteration are the strongly optimal pairs $\langle m_{\texttt{sbna}}, t_{\exists\neg\forall}\rangle$ and $\langle m_{\texttt{all}}, t_{\forall}\rangle$ and this leads to the blocking of all pairs with the form $m_{\texttt{some}}$:



The BIOT-algorithm terminates here and leaves the form $m_{\texttt{some}}$ dangling.

The IBR model, on the other hand, of course replicates the predictions of bidirectional optimality for messages $m_{\texttt{sbna}}$ and $m_{\texttt{all}}$. But in contrast to the

13. In this diagram and the following diagrams, only strict preferences are drawn. Form-meaning pairs that are not in the generator are left blank in order to indicate a difference between blocked and non-generated form-meaning pairs. Recall that the generator translates into semantic meaning in the corresponding signaling game in which the sender cannot send false signals.

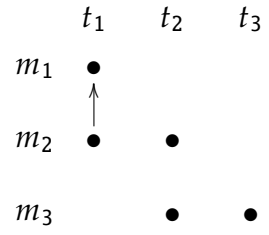
BIOT-algorithm, IBR does *not* block the interpretation $t_{\exists\neg\forall}$ for later association with the form $m_{\texttt{some}}$. With FI assumption, the IBR model does not terminate here, but evolves into a prediction different from weak optimality, as we have seen in section 2.3. Abusing the OT-diagrams to represent the prediction of the IBR model succinctly for visual comparison, here is the fixed point of the IBR model with FI assumption:

	$t_{\exists\neg\forall}$	$t_{\forall}$
$m_{\texttt{sbna}}$	$\odot$	
$m_{\texttt{some}}$	$\odot$	$\times$
$m_{\texttt{all}}$		$\odot$

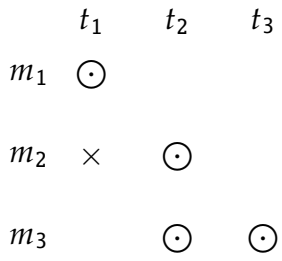
Although this prediction does depend on weak k-dominance, it does not depend on whether the receiver is sophisticated or not. The important point about this example is that the IBR model does not necessarily replicate the strong blocking behavior inherent in weak optimality: in BIOT, if a form or a meaning is blocked, it will never be revived, but not so in IBR. The example furthermore shows that this strong blocking behavior leads to unintuitive predictions for scalar reasoning, if we include non-ambiguous, yet nominally more costly forms; in other words, BIOT does not seem to include enough forward induction reasoning to overcome the symmetry problem.

The suspicion may be raised that IBR differs from BIOT here only because of forward induction, and that the basic IBR model without FI assumption coincides with the predictions of the BIOT-algorithm and hence, in the limit, with weak optimality. This is not so, as another example demonstrates. Even with naïve updaters, and absolutely unspectacular flat prior, cheap-talk interpretation games, IBR does not follow the conservativeness of the BIOT-algorithm.

IBR Overrules Optimality. The next example shows that the IBR model may give up associations between a form and a meaning that were labeled optimal by the BIOT-algorithm in a previous round. Suppose there are three forms and three meanings with semantic meaning as indicated by bullets in the following diagram.



The arrow represents the only *strict* preference between form-meaning pairs according to the Blutner ordering, if we assume that all forms are equally costly and all meanings are initially equiprobable.[14] The BIOT-algorithm will return the following output after one round of iteration after which it also terminates:



The IBR model, in contrast, yields the same prediction for an unsophisticated R_2 in the corresponding interpretation game, but has not yet reached a fixed point. With R_2's interpretation behavior as in the above diagram, S_3 will send only m_2 in state t_2, because this gives her an expected utility of 1 instead of an expected utility of $1/2$ for sending m_3. This is what R_4 realizes and the fixed point is reached with IBR's prediction as in the following diagram:

14. Although it is not strictly necessary to back up the example with reasonable content for the structural point that I would like to make, we can think of this as a scalar implicature case between the three forms m_1 for "it is certain that p", m_2 for "it is likely that p", and m_3 for "there is a remote chance that p." For the sake of the example, let's take the following test as indicative for the semantics chosen above:

(i) a. It's likely that p, and it's maybe even certain that p.

b. There is a remote chance that p, and it's maybe even likely that p.

c. ? There is a remote chance that p, and it's maybe even certain that p.

	t_1	t_2	t_3
m_1	$\odot$		
m_2	$\times$	$\odot$	
m_3		$\times$	$\odot$

That means that, effectively, IBR is not stuck with the strongly optimal pair $\langle m_3, t_2 \rangle$, but weak optimality is. The example shows how IBR is not committed to monotonicity of optimality in establishing form-meaning associations.

Weak Optimality from Myopic Update. These last two examples suggest that the conservativeness of the BIOT-algorithm in terms of blocking and optimality can be modelled in IBR terms only if we assume that the receiver forms posterior beliefs in such a way that all form-meaning associations of the BIOT-algorithm are respected. Towards this purpose, say that a receiver of strategic level $2i$, $i \geq 1$, is MYOPIC if he computes his posterior as

$$\mu_{2i}(t|m) = \Pr(t|\mathrm{Opt}_i(m))$$

whenever $\mathrm{Opt}_i(m)$ is non-empty. A myopic R_{2i} cares for nothing else than the optimal form-meaning associations at round i when he hears a message that is part of an optimal form-meaning pair. With this strong assumption about receiver belief formation we can show that, somewhat unsurprisingly for sure, weak optimality in an OT-system whose ordering defines the expected utilities of R_0 and S_1 corresponds to the limit behavior of a myopic receiver in an interpretation game.

Proposition 4.3.2. If for some message $\mathrm{Opt}_i(m) \neq \emptyset$, then $\mathrm{Opt}_i(m) = R_{2i}(m)$ for myopic receivers.

Proof. Let $\mathrm{Opt}_i(m) \neq \emptyset$. This means that, in an interpretation game, a myopic receiver of level $2i$ plays:

$$R_{2i}(m) = \max_{t \in T} \Pr(t|\mathrm{Opt}_i(m)).$$

We therefore need to show that all states in $\mathrm{Opt}_i(m)$ are equally likely *a priori*. Suppose this is not so, i.e., let there be t and t' in $\mathrm{Opt}_i(m)$ for which $\Pr(t) > \Pr(t')$. In that case, it is impossible for $\langle m, t' \rangle$ to be in Opt_i, because this pair is blocked by the (optimal) $\langle m, t \rangle$. It follows that $\mathrm{Opt}_i(m) = R_{2i}(m)$. □

Reflection. A couple of remarks on this last result are necessary. First of all, this epistemic interpretation of weak optimality seems very much like a brute-force result. Myopic update is certainly a very strong, seemingly artificial assumption about belief formation as it basically requires the formation of posterior beliefs after each optimal form to neglect everything except that form's optimal meanings. Still, it seems that this really is the way we *should* characterize weak optimality in terms of restrictions on belief formation in interpretation games. The strictly monotonic removal of blocked and optimal pairs in the biot-algorithm does not square with interpretation in ibr *at all*: in a manner of speaking, ibr reconsiders *all* possible form-meaning pairs in higher rounds of iteration. So, in order to restrict ibr not to consider certain possibilities, some drastic undermining of ibr's reasoning mechanism is necessary. Consequently, it is not the positive result *that* ibr can model interpretation behavior based on weak optimality that is of importance here, but rather *at what expense* we can characterize weak optimality, i.e., how much of ibr we apparently have to give up in order to assimilate weak optimality.

But, of course, the above is only a *sufficient* characterization in the sense that it gives sufficient epistemic conditions for interpretation based on weak optimality. It thus suffers the same fate that, e.g., epistemic characterization results of game theoretic solution concepts face: it is hard to argue conclusively that a given sufficient characterization is also necessary in the sense that other possible sufficient characterizations are less systematic, elegant or plausible, and that other conceivable systematic, elegant and plausible characterizations are not sufficient. Consequently, I cannot claim that myopic update is a necessary or even the best characterization of weak optimality. There is clearly still room for improvement in future research here.

Finally, if we accept my characterization, it could be argued *for* myopic update and against ibr that myopic update is, in a natural sense, much more resource efficient, and therefore preferable: myopic update could be regarded as a 'fast and frugal' heuristic that lumps together forms and meanings, and that would just not let go of previous associations for the sake of simplicity of calculation. I would certainly be very much in favor of an epistemic interpretation of weak optimality in terms of an efficient heuristic if such a heuristic was defensible. But I do not see any pressing conceptual justification for exactly *this* heuristic. It is moreover not the case that the heuristic in question excels by superior empirical predictions that other approaches could not reach. I suggest that in the absence of reasons to the contrary we should stick to sophisticated updating as the normative standard.

Summary. To give a résumé of this chapter so far, I have argued against the widely accepted idea of Dekker and van Rooij (2000) that bidirectional optimality should be linked to Nash equilibria in strategic games. This, my argument went, is simply not the way we conceive of sets of optimal form-meaning pairs, which intuitively specify conditional, not unconditional, production or interpretation behavior. Instead, I proposed to relate BiOT to signaling games, in particular by translating an OT-system into an interpretation game such that its ordering gives the expected utilities of the lowest optimizing sender and receiver types in the IBR model. This resulted in a straightforward characterization of unidirectional optimality as Bayesian rational behavior and of strong optimality as the interpretation of an unsophisticated level-2 receiver. The characterization of weak optimality in terms of myopic receivers may have seemed somewhat forced. But still, in order to pair IBR reasoning, which is fairly liberal in its association of forms and meanings, with the BiOT-algorithm, which was shown to be fairly conservative, a strong assumption about locally incremental, 'once-and-for-all' belief formation seemed necessary.

These formal results also shed light on the issue of how OT's optimality notions should be interpreted. I have in particular addressed the view that bidirectional optimization requires interlocutors to take the perspective of their conversational partners. It is this that the parallelism with IBR made more precise. If what I suggested in this chapter is correct, then we should think of optimality notions in pragmatic applications as centered on a production perspective, in which basic optimizing for production already subsumes unsophisiticated interpretation behavior. Strong optimality in interpretation then comes forward as the result of taking this production perspective as the basis for interpretation.[15] Weak optimality, on the other hand, presents itself as a more challenging interpretation process which takes further steps of iteration into account. Linking BiOT and IBR in the way suggested, the results of this chapter support the position of Hendriks et al. (2007) who write:

> "We can view bidirectional optimization as a mechanism describing human linguistic competence while acknowledging that the recursion allowed for by this mechanism is limited by performance factors."
>
> (Hendriks et al. 2007, chapter 5 section 5.6.1)

This much does not say anything yet about an evolutionary interpretation of OT, in particular of weak optimality. So when Blutner and Zeevat (2008)

15. Strong optimality thus not only lines up formally but also conceptually with the optimal assertions approach of Benz (2006) and Benz and van Rooij (2007) (see also Franke 2008*a*).

write (in their footnote 11) that "the solution concept of weak bidirection can be seen as a rough first approximation to the more adequate solution concepts of evolutionary game theory that describe the results of language change" especially a diachronic interpretation of IBR seems again like the closest game theoretic counterpart. The last two examples of this chapter that demonstrated the difference between the BIOT-algorithm and IBR also suggest that indeed the latter could turn out the better, more refined evolutionary mechanism to select for optimal communication. Although I tend to believe that the IBR model as presented here is equally suited as a diachronic model, I prefer to postpone a more careful examination of this to another occasion.

4.4 Scalar Implicatures in Language Acquisition

BIOT seeks to explain asymmetries in language acquisition by appeal to the relative difficulty for young children to optimize bidirectionally. It is thus interesting to have a closer look at particularly the developmental pattern in the acquisition of scalar implicatures. Is there a production/comprehension mismatch or any other peculiar asymmetry in the acquisition of the ability to handle scalar implicatures? And if so, does BIOT or IBR help explain it?

4.4.1 Overview of Some Recent Studies

LOGICAL CHILDREN. Ira Noveck's was the first in a fairly recent series of thorough investigations into children's ability to compute (scalar) implicatures and the developmental course of acquisition of pragmatic competence (Noveck 2001). Noveck's study was designed —building on early work by Smith (1980)— as a JUDGEMENT TASK where subjects had to "agree" or "disagree" with sentences presented to them. Critical to an assessment of scalar inference in subjects were sentences like (61).

(61) Some giraffes have long necks.

Acceptance of such sentences was counted as what I will call the SEMANTIC RESPONSE (in this task): subjects accepting a sentence like (61) apparently did not compute the scalar implicature associated with the quantifier "some", because if they had, they should have rejected the sentence based on world knowledge. Rejection of such sentences, on the other hand, was counted as what I will call the PRAGMATIC RESPONSE, for it was taken to indicate the ability to compute and integrate the scalar implicature into the overall meaning of

the to-be-evaluated sentence.[16]

Noveck tested children aged 7 to 11 in this judgement task, and compared performance with adult controls. In a nutshell, his data showed that children gave significantly fewer pragmatic responses than adults, although they performed at adult level in control sentences testing for world knowledge and linguistic competence. In slightly more detail, 89% and 85% of the children aged 7-8 and 10-11 respectively did *not* reject critical sentences like (61), thus showing the SEMANTIC RESPONSE in this task. This is in contrast to 59% of adults who gave the pragmatic response. Noveck's study thus suggests, among other things, that children are —as he put it— more logical than adults: pragmatic competence in comprehension seems to develop late.

EVIDENCE FOR EARLY PRAGMATIC COMPETENCE. Still, a strong conclusion to the extent that children are incapable of pragmatic reasoning in general, or of computing scalar implicatures in particular, is *not* warranted. In a follow-up study Papafragou and Musolino (2003) elicited far more pragmatic responses than Noveck in even younger children of around 5 years of age. Their study consisted of two experiments which in conjunction support the view that (i) Noveck's results were correct in that pragmatic competence takes time to develop, but that (ii) even very young children are capable of pragmatic responses —though not at adult level— if the task is amended adequately.

In Papafragou and Musolino's first experiment subjects had to evaluate the answer of a puppet figure to a question about a previously acted out scene. For example, in the implicature-critical condition subjects would observe a group of toy horses all jumping over a fence; the puppet figure, who observed the scene along with the subject, was then asked what had happened and she would reply —in the critical condition involving quantifier "some"— that some of the horses jumped over the fence; subsequently the subject was asked whether the puppet "answered well" or not. The semantic response in the critical condition of this judgement task is to say that the puppet answered well, while the pragmatic response is to say that the puppet did not answer well. Papafragou and Musolino tested 30 children, aged 4;11 to 5;11, and 30 adult subjects on the contrast between

(i) quantifiers "all" and "some";

(ii) numerals "three" and "two";

16. There are several points of criticism to raise against this interpretation of subjects' responses in Noveck's task. We will come to this below.

	Adults	Children
all/some	92.5%	12.5%
three/two	100%	65%
finish/start	92.5%	10%

Figure 4.3: Percentage of pragmatic responses in Papafragou and Musolino's first experiment

(iii) verbs "finish" and "start."

The reported pragmatic responses of both children and adults in this experiment are given in figure 4.3. For the present discussion the most interesting result is that 92.5% of adult subjects gave pragmatic responses in the some-all contrast, while only 12.5% of 5-year-olds did. This confirms Noveck's previous conclusion that children do not draw scalar inferences at the same rate as adults do.

However, Papafragou and Musolino's second experiment qualifies this conclusion. The set-up in their second experiment was the same as in the first, except that in the second experiment

1. there was a main character in the acted-out scene who was faced with a challenge, such as catching all of the horses;

2. the puppet figure then commented on the success of this main character in meeting the challenge;

3. subjects were told that the puppet sometimes would say something "silly" and that the subject should help the puppet "say it better";

4. subjects were trained to correct the puppet on pragmatic anomalies in a previous naming task of the same pattern.

It seems fair to say that the second experiment made the task clearer to the subjects by raising the relevance of a pragmatically correct statement: the puppet wants to learn how to speak well, and the question whether the main character achieved his task foregrounded the distinction between, e.g., elements "some" or "all." Indeed, under these conditions, Papafragou and Musolino found that the 5-year-old subjects' performance on critical "some"-sentences went up to 52.5% of pragmatic responses (90% for numeral "two" and 47.5% for "start"). This is still not adult-like performance, but shows that the nature

of the task has a clear influence on eliciting pragmatic responses. In sum, a careful conclusion about children's ability to compute scalar implicatures based on Papafragou and Musolino's results is that young children overall do not respond as pragmatically as adults, but that pragmatic responses in children can be facilitated by different task designs, in particular if the relevance of responding pragmatically is highlighted by design and training.

Pragmatic Competence in an Action-Based Task. Pouscoulous et al. (2007) took Papafragou and Musolino's approach even further. The group hypothesized that pragmatic inferences come at a cost and that therefore pragmatic responses could be elicited from even the youngest subjects proportional to the simplicity of the task. To test their hypothesis, Pouscoulous et al. not only replicated Noveck's judgement task, but they also set up an action-based task which was predicted to further facilitate pragmatic responses in young children. Here is Pouscoulous et al.'s experimental set-up in some detail.

Subjects were presented with five boxes and five tokens that were arranged in one out of three possible scenarios in front of them:

1. in the *subset scenario* two boxes contained exactly one token and three boxes were empty;

2. in the *'all' scenario* all five boxes contained exactly one token; and

3. in the *'none' scenario* none of the five boxes contained a token.

In these scenarios subjects heard the sentences in (62). These were presented as a puppet figure's wish with which the subjects were asked to comply.

(62) a. I would like all the boxes to contain a token.
b. I would like some boxes to contain a token.
c. I would like no box to contain a token.
d. I would like some boxes to contain no token.

Subjects were free to add tokens to boxes, remove tokens from boxes or leave everything as is. Consequently, there are two critical conditions in this task where we can distinguish semantic and pragmatic responses. For one, if a subject removed tokens from a box in the 'all' scenario when hearing the sentence (62b), this would count as a pragmatic response, whereas if no token was removed in this condition this would count as a semantic response. Similarly, if a subject added a token in the 'none' scenario when hearing the

	Adults	7 years	5 years	4 years
'all' scenario/some utterance	14%	17%	27%	32%
'none' scenario/some-not utterance	14%	41%	30%	41%

Figure 4.4: Percentage of semantic responses in Pouscoulous et al.'s second experiment

sentence (62d), this would count as a pragmatic response, whereas if no token was added, this would count as a semantic response.

Pouscoulous et al. tracked the responses of three groups of children with mean age 4;5, 5;6 and 7;5 respectively and compared performance to adult controls. Children's performance in non-critical conditions was adult-like.[17] But in the two critical conditions children showed more semantic responses than adults, just as in previous experiments. The developmental pattern across groups for the critical conditions is given in figure 4.4 which lists the percentage of semantic responses.

Results for the first critical condition —the 'all' scenario in connection with the sentence (62b)— show a monotonic rise in pragmatic responses across age groups, suggesting that children grow gradually towards pragmatic maturity. The second critical condition —the 'none' scenario in connection with the sentence (62d)— was less clear in this respect and, Pouscoulous et al. reason, reflects the general difficulty of processing negated statements (see the paper for in-depth discussion). Noteworthy, in any case, are especially two facts about the first critical condition, namely that (i) a simpler action-based task elicited more pragmatic responses from young children than reported in previous studies, yet (ii) still 32% of 4-year-olds did not manipulate the 'all' scenario when confronted with the puppets wish in (62b), despite otherwise apparently comprehending perfectly well the meaning of quantifiers "all" and "some." The hypothesis that Pouscoulous et al. (2007) started out with seems to hold: since pragmatic inference is costly, pragmatic performance is proportional to age —due to a steady increase in computational resources— and anti-proportional to task complexity.

17. A caveat applies here, because whether, for instance, adding exactly one token to a box in the subset scenario when hearing sentence (62b) is an illogical response and should be classified differently from not manipulating the boxes in this case, may make for different conclusions (see Pouscoulous et al. 2007, for discussion).

Interim Conclusions & Reflection. So far, we have reviewed experimental evidence that the pragmatic ability to compute scalar implicatures generally seems to be acquired later than adjacent semantic competence. Still, the proportion of pragmatic responses by young children varied in the studies that we have looked at so far. Pouscoulous et al.'s study gathered more pragmatic responses from younger children, and this is plausibly so because their task was simpler than those of Noveck and Papafragou and Musolino. Minor differences notwithstanding, the main difference between studies was certainly the nature of the task: judgement vs. action-based. The results so far suggest that action-based tasks indeed elicit more pragmatic responses, but a more direct comparison of performance under both paradigms would clearly be welcome to shed more light on this issue.

However, the common-sense explanation that action-based tasks are simpler, in that they require less cognitive resources, though certainly appealing, raises the question what exactly makes a judgement task more resource-intensive and therefore difficult for young children. Also, while it may be rather uncontroversial that Pouscoulous et al.'s action-based task tests for subjects' pragmatic comprehension skills, we should inquire more carefully into the nature of the judgement task, in particular the variety applied by Papafragou and Musolino: what exactly are we testing when we ask whether an utterance of a puppet figure is "acceptable"? Bluntly put, are we testing for comprehension or production here? On the face of it, perhaps, this judgement task seems to test on competence in production, since, after all, subjects are asked whether the puppet *said* it correctly. If that is so, the difference in performance between action-based and judgement tasks would suggest a comprehension/production mismatch, where production lags behind comprehension, which in turn could be explained in terms of the natural resource-intensity of production over comprehension.

But it does not seem quite that straightforward to say what Papafragou and Musolino's judgement task assesses. Take, for instance, a critical underinformative utterance that "some X are Y" when in fact all X are Y. On the one hand, a pragmatic response in this case may be taken as evidence for *production competence*, because subjects may be assumed to reject the utterance if and only if they themselves would not use it in the presented circumstances. But, on the other hand, rejection of the target utterance may also be taken as indicative of *comprehension competence*, because subjects may be assumed to reject the target if and only if they compute the associated scalar implicature which makes the utterance infelicitous. Of course, if we were to assume

that young children's production and comprehension are just mirror images of one another, the two perspectives would collapse into one and a pragmatic response would be indicative of a combined production/comprehension competence. But, as section 4.3.1 showed, there are many cases of mismatch between production and comprehension in language acquisition. It does not seem possible to tell analytically whether pragmatic responses in the critical conditions in Papafragou and Musolino's judgement task are evidence for pragmatic comprehension, production or maybe even something beyond that. To test this and to complete the picture unequivocally, a more direct way of probing either skill in the laboratory is needed.

Comparing Production, Comprehension and Judgements. These considerations lead to the wish for a study that combines and compares directly (i) a task which clearly tests for comprehension competence, with (ii) a judgement-based task, and (iii) a task which tests unambiguously for production competence. Such data is presented by Katsos and Bishop (2009), an early glimpse of which was presented at esslli 2008 in Hamburg (Breheny and Katsos 2008). Katsos and Bishop conducted a developmental study with children of four age groups (5, 7, 9 and 11 years of age) in which they combined a judgement-based task, as used by Papafragou and Musolino, together with a production task and a picture-matching task.

The production task of Katsos and Bishop resembled the judgement task in that the subjects observed a scene together with a puppet character. But, unlike in the judgement task, the puppet now stated that it did not know how to describe the scene, and so the subjects were asked to help out and give a description on behalf of the puppet. This task then unambiguously tests for subjects' production competence.

The picture-matching task of Katsos and Bishop had subjects choose two pictures of situations in which, e.g., a mouse had either picked up only some, or all of the carrots that it was intended to pick up. The subjects were then asked, in the critical condition, which picture fitted a description such as "the mouse picked up some of the carrots." This task clearly tests for subjects' comprehension competence.

Katsos and Bishop's results show that 5-year-olds give overwhelmingly correct responses in semantic controls, as well as pragmatic responses in the production task and the picture-matching task. Still, 5-year-olds fail on a large scale to give pragmatic responses in the judgement task. These subjects readily *accept* an underinformative utterance in the judgement task *although*

they would not be underinformative in the production task, and would grasp the corresponding pragmatic inference in the comprehension task.

The developmental part of Katsos and Bishop's study moreover seems to show that for all age groups the ability to give pragmatic responses in the production and comprehension tasks soon matches performance in semantic controls, whereas the percentage of pragmatic responses in the judgement task only gradually rises to match performance in semantic controls as subjects mature. These data therefore suggest that children are competent hearers and speakers when it comes to scalar implicatures, but are not competent judges of other speakers' (production) competence. In other words, it seems that the ability to reject pragmatically infelicitous statements takes more time to develop than comprehension and production competence as such.

4.4.2 Tolerance vs. Conceptualization

There are several conceivable explanations for this pattern. Certainly, the hypothesis of Pouscoulous et al. that processing cost plays a role is very much compatible with these extended findings. Both the production task as well as the picture-matching task are in an intuitive sense *easier* than the judgement task. Task complexity, and proper task understanding by young subjects, most definitely plays a role in the success of showing pragmatic performances. But an explanation in terms of processing costs and task complexity is also, in some sense, not entirely satisfactory, because it leaves open what exactly the added complexity of a judgement task is.

Pragmatic Tolerance. An alternative explanation of the delayed acquisition of pragmatic competence in judgement tasks is to assume that age is, so to speak, anti-proportional to pragmatic charity: the younger a child the more tolerant it is towards pragmatic infelicity, while nonetheless objecting strongly to semantically false statements. This pragmatic tolerance hypothesis is proposed as a possible explanation by Katsos (2008*a*) and Katsos and Bishop (2009). The pragmatic tolerance hypothesis explains the discrepancy between performance in judgment tasks as compared to other tasks: by assuming that children do not hold speakers responsible for pragmatic infelicity and thus accept underinformative statements.

What possibly speaks against this pragmatic tolerance hypothesis is recent data accumulated in support of the so-called *Question Answer Requirement* hypothesis about children's interpretation of scopally ambiguous sentences

(Hulsey et al. 2004). The details of the debate about models of children's scope disambiguation are inessential, but it does add to the present concern to note that even very young children of age 3–5 were found to reject an utterance of a sentence like (63) in a judgement task in situations where there was a true interpretation available that reflected surface scope, but which did not address the relevant question under discussion (see Gualmini 2007, 2008, for details).

(63) Some of the pizzas were not delivered.

In other words, the surface scope reading would have allowed the children to accept the sentence, but still they rejected it — as comments showed: on the basis of the inverse scope reading. This is evidence against the idea that young children are more charitable than interested in relevance, at least when it comes to the resolution of syntactic ambiguity. Of course, pragmatic tolerance does not imply lenient syntactic disambiguation, but the case makes clear how modular —and to my mind therefore implausible— an assumption pragmatic lenience is.

The View from Normativity. I would then like to tentatively advance an alternative explanation of the data by suggesting that children are *not* tolerant in the sense that they are equipped with full-fledged pragmatic capabilities which they then do not demand to be displayed by others, due to their young, inexperienced and forgiving nature. Rather, I would like to propose that young children may have just enough pragmatic capabilities to succeed in action-based, production and picture matching tasks, but not enough to succeed in judgement tasks. To see what is at stake, let's have another close look at the logic behind the judgment task.

Judgement tasks, I would like to argue, might require strictly more pragmatic maturity than is necessary for linguistic behavior that demonstrates pragmatic comprehension and production proficiency. Take, as before, the critical underinformative utterance "some X are Y" in a situation where all X are Y. Even if subjects would not use the underinformative sentence themselves, this does not necessarily mean that they would reject *someone else's* utterance. Surely, perhaps subjects are more forgiving or tolerant in judging another's productive performance. Or, maybe, children would not be tolerant *at all*, would they not lack the necessary concept of normativity here: perhaps young subjects merely lack the full introspective power to justify what they are doing right without knowing why they are and why everybody else

should do so too. This means that it is not necessarily the case that pragmatic production competence —in a, say, 'behavioral sense'— implies rejection in a judgement task.

And a similar argument applies to comprehension. Even if subjects do not reject a critical target utterance, this does not mean that they necessarily take the target merely at semantic value. It is perfectly conceivable that subjects do interpret pragmatically —taking a "some"-statement to refer to a "some but not all"-situation— without at the same time transcending their own interpretation behavior as a basis for a normative judgement. To wit, a child may arrive at the pragmatically correct interpretation of "some" without conceptualizing that this is what it and everybody else is *and should be* doing. And, again, this means that it is also not necessarily the case that pragmatic comprehension —in a similar 'behavioral sense' of the term— implies rejection in a judgement task.

According to this explanation it is not that children are tolerant as such, but they behave tolerantly because they lack the conceptualization of the pragmatic norm necessary to assuredly reject underinformative utterances. This CONCEPTUALIZATION HYPOTHESIS is the alternative explanation that I would like to suggest. We may think of the conceptualization hypothesis either as an alternative to pragmatic tolerance, or, as I prefer, as a refinement or reduction of it, so as to give an explanation of tolerant behavior.

IBR AND IMPLICATURE ACQUISITION. The conceptualization hypothesis is supported by the IBR model of pragmatic reasoning. First of all, if pragmatic competence is reasoning competence roughly in the sense of the IBR model, then it is most plausible to assume that what develops with age and linguistic experience is the ability to reason deeper, i.e., to advance to higher levels of iterated reasoning. Children of around 4 years of age pass standard first-order false belief tasks, but only two years later will they pass a second-order false belief task (see Wimmer and Perner 1983, and follow-ups). Reasoning about other people's minds takes time to develop, be that because the conceptual skills to do so need to be acquired, or because higher-order ToM reasoning is indeed a resource-intensive operation. In the IBR model sophisticated reasoners not only ascribe a belief (about beliefs about beliefs ...) to their opponents, but also compute their strategies, i.e., form conjectures about rational or optimized behavior. This may add to the complexity of the process and may cause further delays in the acquisition of pragmatic reasoning capabilities. Consequently, it is natural to hypothesize that most young children of age

4-5 are level-1 reasoners —without necessarily knowing, introspectively, that they are, of course— and that only later they will develop into higher level language users.

Interestingly, both level-1 senders and level-1 receivers already display what I have called *scalar implicature behavior*, but this behavior is not yet supported by a fully self-enforcing set of beliefs. Here is the situation in some more detail. In our standard model for scalar reasoning spelled out in section 2.2.2, a level-1 sender will use $m_{\texttt{some}}$ *only* in state $t_{\exists\neg\forall}$. Nonetheless, S_1 believes that her opponent R_0 does not interpret $m_{\texttt{some}}$, we could say, with a scalar implicature. A level-1 sender thus shows scalar implicature behavior without necessarily having *scalar implicature beliefs*. We could therefore say that S_1 shows scalar implicature behavior without having fully conceptualized it, since she does not expect her opponent to show scalar implicature behavior. Similarly, a level-1 receiver will choose $a_{\exists\neg\forall}$ in response to message $m_{\texttt{some}}$, without actually believing that $m_{\texttt{some}}$ is sent only in $t_{\exists\neg\forall}$. Again, R_1 shows scalar implicature behavior even without scalar implicature beliefs, i.e., without the belief that his opponent does so too. That suggests that even level-1 reasoners will give pragmatic responses in action-based, production and picture-matching tasks, because all of these tasks really take the form of a simple signaling game for scalar implicature in which the subjects either take the role of the sender or the receiver.

What about the judgement task then? I would like to suggest that scalar implicature behavior alone, be that comprehension or production competence, is not sufficient for a pragmatic response in the critical condition in a judgement task. What seems minimally necessary for a pragmatic response in a judgement task is that subjects can judge other people's linguistic behavior based on —but crucially on top of— their own pragmatic competence: what is needed is that scalar implicature behavior is also supported by a belief that the opponent shows scalar implicature behavior. In the IBR model this requires at least a sophistication level 2.[18]

18. On a speculative note, moreover, it seems that for a fully developed *normative* stance towards the proper pragmatic use of expressions, even more than sophistication level 2 seems necessary. To be able to say that some expression should be used or interpreted in such and such a way requires one's conjectures about general use and interpretation to be in equilibrium, so to speak: it's how we should do it because it's common expectation that we do so. In other words, it seems that for full normative understanding of pragmatic use, reasoners must have transcended the IBR sequence. — The relation between norms, conventions and mutual or common expectations, however, is a spicy philosophical issue that I will not go into here. And, of course, this is also not needed to account for the acquisition data.

That means that an IBR model of pragmatic reasoning implements the conceptualization hypothesis, making it more precise. This way, the IBR model explains the observed data. It explains why young children show scalar implicature behavior in a simple task testing on proper production and comprehension, and it also explains what exactly is more difficult about a judgement task: a pragmatic response in a judgement task requires higher levels of sophistication in pragmatic reasoning than either the production, the action-based or the picture-matching task.

This is also exactly the reason why the IBR model seems a better formal model to back up and spell out the conceptualization hypothesis than BIOT. If the conceptualization hypothesis is correct, we want differently sophisticated pragmatic competencies to be concisely represented in the model. But, as section 4.3 showed, BIOT's conceptually somewhat underspecified notions do not live up to this challenge. This is a problem, of course, only to the extent that the conceptualization hypothesis stands to further empirical scrutiny. It is thus up to empirical testing to decide between the pragmatic tolerance hypothesis, the conceptualization hypothesis —if these are perceived as mutually exclusive— or any other conceivable interpretation of the rather intricate acquisition data.

Chapter 5

The Pragmatics of Conditionals

"She was amazing. I never met a woman like this before. She showed me to the dressing room. She said: 'If you need anything, I'm Jill.' I was like: 'Oh, my God! I never met a woman before with a *conditional identity*.' [Laughter] 'What if I don't need anything? Who are you?' — 'If you don't need anything, I'm Eugene.' [More laughter]"

(Demetri Martin, *These are jokes*)

ME: [*Commenting on the media reception of Paul Potts' and Susan Boyle's appearance on* TV *show 'Britain's got Talent'*] Obviously, if a person doesn't conform to excessive norms of physical attractiveness, doesn't mean that he or she is untalented or stupid.

SHE: So you are basically saying that beauty doesn't guarantee intelligence. — Wait! Are you trying to tell me I'm dumb?

ME: ?!? (my life)

Chapter Contents

5.1 · Meaning and Use of Conditionals · 226

5.2 · Conditional Perfection · 234

5.3 · Unconditional Readings · 257

Loosely speaking, some conditionals convey more of a conditional meaning than others. We see what is at stake when we compare the by-now classic examples (64) from Geis and Zwicky (1971) and (65) from Austin (1956).

(64) a. If you mow the lawn, I'll give you five dollars.
b. $\rightsquigarrow$ If you don't mow the lawn, I will not give you five dollars.
c. $\not\rightsquigarrow$ I'll give you five dollars.

(65) a. There are biscuits on the sideboard if you want them.
b. $\not\rightsquigarrow$ If you don't want them, there are no biscuits on the sideboard.
c. $\rightsquigarrow$ There are biscuits on the sideboard.

Whereas it is fairly natural to interpret a generic utterance of (64a) to convey also the *obverse* in (64b), the conditional in (65a) does not naturally convey (65b). Rather, (65a) seems to convey that its consequent (65c) is true *unconditionally*, while (64a) certainly does not convey (64c). So, in a superficial manner of speaking, we might say that under their standard readings (64a) expresses a stronger conditional meaning than (65a) does.

The pragmatic strengthening of a conditional like in (64) has been dubbed *conditional perfection* and will be the topic of section 5.2. The conditional in (65a) was eponymous for the class of *biscuit conditionals* that became prototypical examples for conditionals with what I will call *unconditional readings*. Section 5.3 deals with unconditional readings. On the face of it, conditional perfection readings and unconditional readings are very much mirror-image phenomena, and this section consequently aims to show how similar pragmatic mechanisms of contextual enrichment give rise to both of these.

More concretely, the main hypothesis of this chapter is that we can and should explain the bulk of conditional perfection and unconditional readings as 'commonsense inferences': I argue that the correct interpretation of a conditional sentence can often be derived by imposing additional commonsense constraints about the intuitive relatedness of antecedent and consequent on the models that we evaluate the conditional on; which constraints these are has to be defended against common sense on a case by case basis. Only for a few cases of conditional perfection do we need to refer back to genuine pragmatic reasoning about the topic of conversation.

5.1 Meaning and Use of Conditionals

It is not the intention of this chapter to assess or advance complicated semantic theories of conditionals. The following section 5.1.1 only surveys standard

possible-worlds semantics for conditionals, which I take to be, say, sufficiently true for the purposes of the present pragmatic investigation.[1] Section 5.1.2 then introduces a very rough classification scheme of uses that English conditional sentences may have, to the extent that these distinctions are relevant to the subsequent discussion.

5.1.1 Semantics for Conditionals

MATERIAL IMPLICATION. According to a MATERIAL IMPLICATION analysis, the semantic meaning of a conditional $A > C$ is captured by the truth conditions of material implication $\rightarrow$ of propositional logic, as in the following table:[2]

A	C	$A \rightarrow C$
1	1	1
1	0	0
0	1	1
0	0	1

This analysis may seem too simple, because, among other things, it does not appeal very much to our intuitions about negated conditionals. The negation of a conditional as in (66a) intuitively rather means (66b), than (66c), as the negation of a material conditional would predict.

(66) a. It's not the case that if A, then C.

b. If A, then it's (at least) possible that $\overline{C}$.

c. A is true and C is false.

1. Edgington (1995) and Bennett (2003) provide a thorough background on philosophical theorizing about the semantics of conditionals. A neat and concise survey of the semantics of conditionals is given by Kaufmann (2005*b*).

2. The notation $A > C$ refers to a conditional sentence as an *abstract linguistic form*: in other words, the symbol $>$ is not part of any formal language, but is merely an abstract placeholder for different morpho-syntactic ways of conjoining two clauses A and C in a conditional construction. Most of the time, it suffices for our modest purposes here to assume that the clauses A (for antecedent) and C (for consequent) express simple propositions. Often these propositions are taken to denote simple sets of possible worlds, in which case the letters A and C may denote both a linguistic expression *and* a set of possible worlds. I will use notation $\overline{X}$, to denote the negation of proposition X, alongside the more common symbol $\neg$.

Strict Implication. To deal with this issue, we could have recourse to a slightly more elaborate analysis in terms of strict implication. If $\sigma \subseteq W$ is a set of possible worlds, then we say that a conditional $A > C$ is *supported* on σ if all worlds in σ that make A true also make C true. For this analysis, the set σ is a contextually given set of possible worlds that represents either the live options of the common ground, or the information state of a single agent, most often the speaker. Thus conceived, the set σ will usually change dynamically during conversation or under belief update and revision.

To treat conditionals as strict implication suffices for a great number of applications but still there are good arguments for a more refined treatment. To see what is at stake consult your intuition on the pair of sentences in (67).

(67) a. If you strike this match, it will light. $A > C$

b. If you stand in a storm and strike this match, it will not light. $(A \wedge R) > \overline{C}$

It seems defensible that both conditionals can in fact be true at the same time, for there is no inherent contradiction in either of the following statements:[3]

(68) a. If you strike this match, it will light, but if you stand in a storm and strike this match, it will not light. $(A > C) \wedge ((A \wedge R) > \overline{C})$

b. If you stand in a storm and strike this match, it will not light, but if you strike this match, it will light. $((A \wedge R) > C) \wedge (A > C)$

A similar argument applies to *counterfactual conditionals*: again, it is certainly possible for both of the sentences in (69) to be true simultaneously (cf. Goodman 1947; Lewis 1973).

(69) a. If you had struck this match, it would have lit. $A > C$

b. If you had been standing in a storm and struck this match, it would not have lit. $(A \wedge R) > \overline{C}$

But then the problem for a strict implication analysis is that there is no information state σ except the trivial *absurd state* $\sigma = \emptyset$ which supports both statements in (67), respectively (69).

3. Indeed, the order of presentation of the indicative conditionals in (68), as well as the counterfactuals in (69), matters in discourse (cf. Veltman 1985; von Fintel 2001*b*; de Jager 2009): a so-called *Sobel sequence*, as in (68a), sounds more felicitous than a so-called *reverse Sobel sequence*, as in (68b). But this is not important for our present concerns.

Order-Sensitive Implication. The standard solution to this problem is to additionally include into the semantics a comparative notion on the set of possible worlds that conditionals are evaluated on. An order-sensitive implication analysis has a conditional $A > C$ evaluated as either true or false in a world w with respect to a suitable modal structure $\langle R_w, \preceq_w \rangle$, where $R_w \subseteq W$ is the set of worlds accessible from w and $\preceq_w$ is a well-founded ordering on R_w.[4] Using the ordering information in such a modal structure we say that a conditional $A > C$ is true in w iff C is true in all the $\preceq_w$-minimal worlds in R_w in which A is true. Formally, define

$$\text{Min}_w(A) = \{v \in R_w \cap A \mid \neg\exists v' \in R_w \cap A \ : \ v' \prec_w v\}$$

as the set of $\preceq_w$-minimal A-worlds in R_w and define:

$$A > C \text{ is true in } w \quad \text{iff} \quad \text{Min}_w(A) \subseteq C. \tag{5.1}$$

This analysis indeed allows the pair in (67) to be true at the same time, due to the additional ordering of possible worlds: clearly, there are sets R_w and orderings $\preceq_w$ such that the *minimal* worlds in which the match is struck are worlds with good weather conditions for the match to light when struck; still, the minimal worlds where the match is struck *and* there is a storm may be worlds where the match does *not* light. Similarly, of course, for (69).

Notice that order-sensitive implication is basically an abstraction over several possible semantics for conditionals, as long as we are vague about the conceptual interpretation and the formal properties of R_w and $\preceq_w$. Indeed, different kinds of conditionals will require slightly different conceptual interpretations and also different formal properties (see section 5.1.2). If we adopt different specific assumptions, we obtain (close) equivalents of different semantics for conditionals (e.g. Stalnaker 1968; Lewis 1973; Veltman 1986; Kratzer 1991). Since this chapter is mainly about the pragmatics of conditionals, I will try to remain as uncommitted and general as possible here.

Conditionals and Modals. Conditionals are closely related to modals. Some authors have argued that the antecedents of conditionals should be analyzed as restricting the domain of quantification of a (possibly implicit) modal in the consequent (Lewis 1975; Kratzer 1991). Indeed, there is a clear distinction in meaning between the two conditionals in (67a) and (70).

4. The order $\preceq_w$ is well-founded on R_w iff for all $X \subseteq R_w$ there is a $\preceq_w$-minimal element in X. This *limit assumption* (Lewis 1973) is adopted here only for ease of formalization.

(67a) If you strike this match, it will light.

(70) If you strike this match, it might light.

This difference is not visible when I write $A > C$ as a general stand-in for a conditional sentence. To make this distinction visible where it matters I will write $A \;\Box\!\!\Rightarrow\; C$ for sentences like (67a) with a universal modal in the consequent, and $A \;\Diamond\!\!\Rightarrow\; C$ for sentences like (70) with an existential modal in the consequent. The former should indeed be analyzed as in (5.1), the latter should be analyzed as follows:

$$A \;\Diamond\!\!\Rightarrow\; C \text{ is true in } w \quad \text{iff} \quad \mathrm{Min}_w(A) \cap C \neq \emptyset. \tag{5.2}$$

Order-sensitive implication actually incorporates the idea of modal domain restriction by the antecedents of conditionals. Given a world w with accessible worlds R_w and an ordering $\preceq_w$ that capture the relevant modality, we say that a universal modal statement $\Box C$ is true in w iff all $\preceq_w$-minimal worlds in R_w make C true. Define

$$\mathrm{Min}_w = \left\{ v \in R_w \mid \neg\exists v' \in R_w \;:\; v' \prec_w v \right\}$$

as the set of $\preceq_w$-minimal worlds in R_w. With this define the truth of a universal modal statement as follows:

$$\Box C \text{ is true in } w \quad \text{iff} \quad \mathrm{Min}_w \subseteq C. \tag{5.3}$$

And, similarly, for existential modals:

$$\Diamond C \text{ is true in } w \quad \text{iff} \quad \mathrm{Min}_w \cap C \neq \emptyset. \tag{5.4}$$

Looking at things this way we find that if $\mathrm{Min}_w \cap A \neq \emptyset$, then the semantics in terms of order-sensitive implication comes down to a semantics of the modals $\Box C$ and $\Diamond C$ after the domain of quantification R_w for the modal has been restricted to worlds where the antecedent is true.

Belief Dynamics and Ramsey Test. If, on the other hand, $\mathrm{Min}_w \cap A = \emptyset$, then the antecedent of a conditional may be said to shift the context of interpretation of the modalized consequent. This is related to another very prominent idea about the procedural interpretation of conditionals. Ramsey (1931) suggested in passing that conditionals are to be evaluated in a three-step procedure. In the words of Robert Stalnaker the so-called *Ramsey test* takes the following form:

> "First add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the belief in the antecedent); finally, consider whether or not the consequent is then true." (Stalnaker 1968, p. 102)

We find the Ramsey test in our general semantics if we think of the modal structure $\langle R_w, \preceq_w \rangle$ as an agent's actual beliefs and her belief revision policies. Let the $\preceq_w$-minimal worlds in R_w be the set of worlds an agent actually holds possible. Moving to the $\preceq_w$-minimal worlds where A is true comes down to hypothetical belief change: hypothetically adopting the belief that A or dropping the belief $\overline{A}$. In this hypothetical belief state, the agent then checks whether she believes that C is true, i.e., she checks whether $\Box C$ holds, or, she checks whether she considers C possible, i.e., whether $\Diamond C$ holds. In this sense, the order-sensitive implication analysis may implement the Ramsey test, as an evaluation procedure of conditionals in terms of belief revision policies, at least for certain interpretations of R_w and $\preceq_w$.

Order-Sensitive Subsumes Strict Implication. If we interpret the modal structure $\langle R_w, \preceq_w \rangle$ as specifying an agent's beliefs and dispositions to revise these beliefs, then strict implication comes out as a special case of order-sensitive implication. We only need to equate σ with the set Min_w. Using this fact, I will at times make use of a strict implication analysis where it eases illustration of an example, despite the fact that the main results are to be derived for the more encompassing notion of order-sensitive implication.

Non-Triviality Presupposition. Order-sensitive implication effectively is a quantification over the set $\mathrm{Min}_w(A)$. It is often desirable to exclude trivial truth of $A \;\Box\!\!\Rightarrow C$ and trivial falsity of $A \;\Diamond\!\!\Rightarrow C$ that arises just because this set is empty. I suggest to do so and speak of a NON-TRIVIALITY PRESUPPOSITION here. Throughout this chapter I will follow common practice and assume that the minimal non-triviality presupposition $R_w \cap A \neq \emptyset$ is always met. Similarly, wherever I resort to strict implication (for ease of explanation) I will make the slightly stronger yet equally common assumption that $\sigma \cap A \neq \emptyset$.

5.1.2 Kinds of Conditionals

With respect to both meaning and use, conditionals are actually a heterogeneous class in which we might want to distinguish different *kinds of conditionals*. Over the years many classification schemes have been proposed in

the literature, all with diverging terminology, and all variously motivated by either mainly functional (Comrie 1986; Sweetser 1990; Dancygier 1998), syntactic (Haegeman 2003) or pragma-semantic concerns (Iatridou 1991; Bhatt and Pancheva 2006). For the purposes of the following discussion I would also like to make a few very rudimentary distinctions that influence the interpretation of our semantics —and possibly the formal properties of the modal structure $\langle R_w, \preceq_w \rangle$— and fix terminology.[5]

Epistemic Conditionals. *Epistemic conditionals* are conditionals like in (71) in which, roughly speaking, the speaker reports on her conditional beliefs concerning propositions that could in principle be known but whose truth the speaker is *subjectively uncertain* of.

(71) a. If Oswald did not shoot Kennedy, someone else did.

b. If the butler hasn't killed her, the gardener must have.

c. If you struck this match (while I wasn't watching), it must have lit.

The antecedents of epistemic conditionals could be analyzed as modifying, possibly implicitly, the *epistemic* modals *must* and *might*. Accordingly, for an analysis in terms of order-sensitive implication the set R_w would contain worlds the speaker (or some other 'supporting' agent) cannot rule out on the basis of some true information or hard evidence. The ordering $\preceq_w$ would then encode doxastic prejudices, hidden assumptions and the like.

Predictive Conditionals. *Predictive conditionals* are conditionals like in (72) in which a prediction is expressed about future courses of events whose occurrence is *objectively* uncertain and therefore cannot in principle be known.

(72) a. If you strike this match, it will light.

b. If Andrea arrives late, Clara will be upset.

c. If it rains tomorrow morning, the barbecue will be cancelled.

The antecedents of predictive conditionals could be analyzed as modifying, possibly implicitly, the future modals *will* and *might*. The set R_w on which a predictive conditional is to be evaluated should be (something like) the set of

5. I don't wish to spend much argument beyond appeal to naïve intuition on a justification of the proposed distinctions. I also do not mean to suggest that this classification is exhaustive and non-overlapping.

future developments of w, and the ordering relation $\preceq_w$ would encode objective and natural assumptions about causality and commonsense expectations about normal courses of events (cf. Morreau 1997; Kaufmann 2005*a*).

Commissive Conditionals. *Commissive Conditionals* are *conditional promises* as in (64a) and (73a), and *conditional threats* as in (73b) and (73c) with which the speaker tries to exert influence on the hearer's decision making.

(64a) If you mow the lawn, I'll give you five dollars.

(73) a. I'll lend you the book, if you lend me your bicycle tomorrow.
b. If you don't stay away from my girl, I'll burn your record collection.
c. If Martha finds out about this, our friendship is over.

Commissive conditionals have consequents that are desirable or undesirable to the addressee. Still more importantly, what sets commissive conditionals off from predictive conditionals is that in the former the consequents refer to actions or events that are under *speaker control* while their antecedents are usually events under hearer control. Here R_w would contain future developments of the world w, as in predictive conditionals, but crucially the ordering $\preceq_w$ should also capture our intuitions about commonsense *social behavior*, i.e., natural dispositions to act by and large *rationally*, in accordance with one's intentions, aims and beliefs.

Counterfactual Conditionals. *Counterfactual conditionals*, or *counterfactuals* for short, are conditionals as in (69) or (74) in which the antecedent has a backward-shifted tense and the consequent is in the subjunctive, usually so as to express counterfactuality in the sense that A is not assumed to be true by the speaker (or only very unlikely, or not endorsed as possible or sufficiently likely at the present stage of the conversation etc.).

(74) a. If kangaroos had no tails, they would topple over. (Lewis 1973)
b. If I were a carpenter and you were a lady, would you marry me anyway? (Tim Hardin)

For an evaluation of counterfactual conditionals, we would interpret the set R_w as *all* conceivable worlds and $\preceq_w$ as a measure of similarity to the actual world w, in terms of facts and what we consider normal courses of events. Usually, it is assumed that $w \in R_w$ and that $\preceq_w$ either satisfies weak centering (see Lewis 1973)

$$w \in \text{Min}_w$$

or strong centering (see Stalnaker 1968):

$$\text{Min}_w = \{w\}\,.$$

5.2 Conditional Perfection

Although they were not the first in history to discuss the phenomenon, it was a short paper by Geis and Zwicky (1971) that subsequently fixed the community's awareness on the interesting observation that conditional sentences such as (75a) are often understood to also convey the reverse direction of fit (75b) and thus to be taken as a biconditional (75c) (for overview see van der Auwera 1997*b*; Horn 2000).

(75) a. If John leans out of that window any further, he'll fall. $A > C$

b. If John does not lean out of that window any further, he will not fall. $\overline{A} > \overline{C}$

c. If and only if John leans out of that window any further, he will fall. $A \Leftrightarrow C$

It is easy to see, however, that (75a) does not imply (75b) under any of the semantic analyses discussed in section 5.1.1 above (although, of course, the attested reading is compatible with all of the analyses considered). Geis and Zwicky consequently hypothesized that this tendency for *conditional perfection* (CP) had a pragmatic default character and formulated a principle that associated conditional surface form with what they called an *invited inference* to perfect conditionals, i.e., to understand them as biconditionals as in (75c). Thus conceived CP is central to Gricean pragmatics, and I would like to investigate in this section what exactly the nature of the attested inference is, under which circumstances it arises and how it can be explained.

But what exactly is a CP-reading? What are we trying to explain? The literature on CP is far from unanimous about this, if the issue is raised at all. Geis and Zwicky (1971) seem to have regarded the strong biconditional readings in (75c) as the CP-reading of (75a). But strictly speaking, this could either be arrived at by an enrichment of (75a) with (75b) or by an enrichment with either of the "only if" conditionals in (76).

(76) a. Only if John leans out of that window any further, will he fall. only if A, C

b. Only if John does not lean out of that window any further, will he not fall. only if $\overline{A}$, $\overline{C}$

In a recent descriptive article, van Canegem-Ardijns and van Belle (2008) distinguish these three possibilities and argue that the ensuing CP-readings differ slightly in strength: a CP-reading as in (75b) is the baseline case and the readings in (76) are both counted as stronger than the baseline. Which reading obtains, van Canegem-Ardijns and van Belle argue, depends on the kind of conditional that is being perfected.[6]

In what follows I will focus on the most general and basic CP-reading which is arguably even weaker than (75b). Take, for instance, the example in (77a).

(77) a. If I bum around, I will miss my deadline. $(A > C)$

b. If I don't, I will not miss my deadline. $(\overline{A} > \overline{C})$

c. If I don't, I might not miss my deadline. $(\overline{A} \diamond\!\!\Rightarrow \overline{C})$

d. If I don't, I will miss my deadline. $(\overline{A} > C)$

Whether the conditional (77a) has a conditional reading or not depends on whether C is true if A does *not* hold. Our semantics allows us to distinguish three basic cases here: (i) the STRONG CP-READING in (77b), (ii) the WEAK CP-READING in (77c), and (iii) the UNCONDITIONAL READING in (77d). Thus conceived, a minimal CP-reading as in (77c) is basically very weak, paraphrasable as saying

(78) $\overline{A}$ is not a sufficient condition for C. $\neg(\overline{A} \boxdot\!\!\Rightarrow C)$

This reading then may or may not be strengthened to a reading that paraphrases as

(79) $\overline{A}$ is a sufficient condition for $\overline{C}$. $\overline{A} \boxdot\!\!\Rightarrow \overline{C}$

Whenever I speak of a CP-reading in the following, I mean *at least* a weak CP-reading. In other words, I take CP-readings to be just the exclusion of an unconditional reading as in (77d), and vice versa.

6. More specifically, van Canegem-Ardijns and van Belle (2008) argue that whereas predictive conditionals do not generally give rise to the stronger CP-readings in (76), commissive conditionals do. In particular, a conditional promise of the form $A > C$ gives rise to the reading that "only if A, C", whereas a conditional threat of the form $A > C$ gives rise to the reading that "only if $\overline{A}$, $\overline{C}$." In other words, if $A > C$ is a conditional promise, we obtain the intuitive reading that C is *not* going to happen, but the hearer may bring it about by A; on the other hand, if $A > C$ is a conditional threat, we obtain the intuitive reading that C *is* going to happen, but the hearer may prevent it by $\overline{A}$.

5.2.1 Approaching Perfection

Before heading into an analysis, it pays to briefly review previous proposals.

Perfection from Lexical Strengthening

Atlas and Levinson (1981), Horn (1984), Levinson (2000) and Horn (2000) treat CP as an I-implicature. The most thorough exposition of this idea is given by Horn (2000) who argues that CP is a case of *lexical strengthening*.[7] The inference from *if* to *iff* should be thought of, Horn suggests, in parallel to other diachronic strengthenings of lexical meanings, such as, for instance, the case of the English *liquor*, specific now for alcoholic beverage, but derived from the more encompassing term *liquid*. In effect, CP is explained as conventionalized pragmatic narrowing of the semantic meaning of the connector *if* and related constructions.

The impression is strong that this is not much of an *interesting* explanation of an interesting phenomenon (see van der Auwera 1997*a*, for similar criticism). In particular, conventionalized lexical strengthening alone cannot easily explain the context-dependence of CP-readings. But it is clear that CP does not arise for certain conditionals, such as (65a), and some conditionals may get a CP-reading in one context, but not in another: I will argue below that CP-readings systematically depend on the topic of conversation and therefore lend themselves to a more detailed and systematic derivation than an account in terms of conventionalized lexical strengthening permits.

Perfection as Scalar Implicature

A more attractive explanation ensues from treating CP as a scalar implicature. Ducrot (1969), Horn (1972), Boër and Lycan (1973), Matsumoto (1995) and van der Auwera (1997*a*), among others, have suggested such accounts. If we want to explain conditional perfection as a scalar implicature, the problem is of course which alternative forms to refer to.

Biconditional and Unconditional as Alternatives. If we assume that the alternatives to be compared are:

(80) $\{A > C, \quad A \Leftrightarrow C\}$

7. Most papers suggesting I-implicature accounts of CP do not explain or derive the attested inference explicitly, but merely list CP as *one* example under many for I-implicatures.

we run straight into the symmetry problem (see sections 2.3.1 and 3.1.2), and are unable to account for CP-readings. A different simple alternative set that promises to be more successful is (81) where we compare the target conditional with the expression "whether A or not, C."[8]

(81) $\{A > C,$ "whether A or $\overline{A}$, C"$\}$

Rawlins (2008*a*,*b*) analyzes these latter expression, which he calls *unconditionals*, as equivalent to the conjunction in (82).

(82) $A > C \ \wedge \ \overline{A} > C$

Under this analysis, it is plain to see that naïve scalar reasoning based on the alternatives in (81) derives CP: if $A > C$ is true and the unconditional as analyzed in (82) is false, then $\overline{A} > C$ must be false; this is equivalent to the weak CP-reading $\overline{A} \diamond\!\!\Leftrightarrow \overline{C}$.

Although the set in (81) derives CP straightforwardly, the problem with this explanation is that it stands in need of justification that the set in (81) is a feasible set of alternatives for Gricean reasoning while the set in (80) is not. To wit, both sets fail all of the common constraints on scalar comparison suggested in the literature (see section 3.1.2). So, despite the fact that negating "whether A or not, C" might give us CP, it is not clear why this should be an alternative expression for scalar reasoning.

Scales from Alternative Antecedents. That, among other things, is why Matsumoto (1995) and van der Auwera (1997*a*) assume different sets of alternatives. The most explicitly spelled-out account is van der Auwera's, who considers the set

(83) $\{A > C,$
$A > C \ \wedge \ B_1 > C,$
$A > C \ \wedge \ B_1 > C \ \wedge \ B_2 > C,$
$A > C \ \wedge \ B_1 > C \ \wedge \ B_2 > C \ \wedge \ B_3 > C,$
$\ldots \ \}$

where B_i are relevant alternative propositions for A. With this, van der Auwera assumes, an utterance of $A > C$ will implicate that $(A \vee B_i) > C$ is false for all B_i. This, together with the truth of $A > C$, implies that $B_i > C$ is false

8. According to van der Auwera (1997*a*), this analysis is due to Ducrot (1969).

for all B_i, so that we get (5.5) as an implicature:

$$\bigwedge_{\{B_i\}} \neg(B_i > C). \tag{5.5}$$

According to van der Auwera, this would then explain CP as a Q-implicature, because the given condition A is the *only* one sufficient for C from the set of alternatives, which yields the CP-reading in (76a).

Is this a convincing explanation of CP that we should adopt? I argue that it is not. First of all, a little nagging. The derivation of (5.5) is not entirely correct under an order-sensitive implication analysis. From a negation of all alternatives to $A > C$ from the set in (83) we can only derive that $B_1 > C$ is false. To arrive at the stronger conclusion in (5.5) we would, strictly speaking, need to assume a different alternative set, namely:

(84) $\{A > C,$
$A > C \ \wedge \ B_1 > C,$
$A > C \ \wedge \ B_2 > C,$
$A > C \ \wedge \ B_3 > C,$
$\ldots \ \}.$

But this is, of course, not a major point of criticism.

Still, even with this alternative set problems continue. It is fairly obvious that the scalar inference in (5.5) does not necessarily entail a weak CP-reading of the kind I am after, at least not for arbitrary alternatives B_i. We need to place additional restrictions on the set $\{B_i\}$ in order to derive CP. One natural and formally sufficient condition is the condition

$$\overline{A} = \bigcup_i B_i. \tag{5.6}$$

To assume (5.6) is to assume that $\overline{A}$ is exhausted by the possibilities B_i, and that the possibilities B_i do not overlap with A. This is not an unnatural additional assumption for a set of alternatives to A, of course. Moreover it helps to formally derive a weak CP-reading as follows: if $\bigcup_i B_i = \overline{A}$ then there is a subset $\mathcal{B} \subseteq \{B_i\}$ such that $\text{Min}_w(\overline{A}) = \bigcup_{B_i \in \mathcal{B}} \text{Min}_w(B_i)$. But then $\text{Min}_w(\overline{A})$ contains some worlds where $\overline{C}$ is true, since every $\text{Min}_w(B_i)$ does.

But if the condition in (5.6) is necessary for a weak CP-reading —necessary, in the sense that it is the most natural condition that is formally sufficient to derive the result— then any instantiation of the scale in (84) that satisfies (5.6) can be abbreviated to

(85) $\{A > C, \quad A > C \wedge \overline{A} > C\}$.

Thus conceived, it transpires that under the most reasonable interpretation of the set of alternatives to A that also works, the set in (84) is equivalent to the set in (81) which compared the conditional with the unconditional "whether A or not, C."

Alternatives and Topics. Does this mean that we have successfully reconstructed the behavior of the problematic set (81) by a different, more plausible, more defensible set (84)? Again, I remain doubtful. The question remains why we should build a set of alternative expressions to $A > C$ by looking at alternatives to the antecedent and not for the consequent (or both). In other words, why is (84) the correct set and not for instance the set in (86)?

(86) $\{A > C,$
$A > C \wedge A > D_1,$
$A > C \wedge A > D_2,$
$A > C \wedge A > D_2,$
$\ldots \}$

Intuitively, the matter may seem obvious: in most standard contexts of utterance of $A > C$, we are rather interested in (an answer to the question after) sufficient conditions for C, rather than (an answer to the question after) the consequences of A. In other words, it seems to be an implicit contextual *topic requirement* that motivates the use of (something like) the scale in (84) in some contexts but not in others. But then we should not think of CP as run-of-the-mill scalar implicature to begin with.

I therefore argue that where CP-readings are to be derived by 'something like' scalar reasoning, the 'something like' is contextual interpretation of the conditional under a given topical question under discussion, and not reasoning about a fixed set of expression alternatives generally associated with a conditional by lexicon or grammar. Still, not all cases of CP require such pragmatic reasoning, but can much more naturally be accounted for by appeal to shared assumptions about natural relatedness of events. This is what I will argue for next.

5.2.2 Two Sources of Perfection

I claim that there are really two distinct sources of CP that also require distinct treatments: (i) CP-readings can arise from shared normality assumptions and

world knowledge of a default kind; (ii) CP-readings can also arise by more tangible pragmatic reasoning about the topic of conversation.

Perfection from Normality

WINTER IN AMSTERDAM. Having lived in Amsterdam for a couple of years, I found it interesting to see how the city administration is prepared to deal with ice and snow in winter time. What surprised me could be expressed perspicuously by the following two conditional sentences:

(87) a. If the canals freeze, the city sends out icebreaker boats to drive through the major canals, but ...

b. ... if it snows, the city does not send out snowplows to drive through the streets.

I am surprised about this, because my hometown has no canals but does have regular snowfall in wintertime. I therefore did not expect to see icebreaker boats on the frozen canals in Amsterdam, but I would have expected to see snowplows at work for traffic safety, since this is a normal occurrence in the town that I grew up in.

Both conditionals in (87) are relevant (to me, and, let us assume, also to the conversation) because they express my surprise, my failed expectations. But what's more important here is that there is also an interesting contrast, as far as common expectations are concerned: I am sure we *all* do *not* expect the city to send out icebreaker boats or snowplows if it does not freeze or snow; more concretely, we all expect that the conditionals in (88) are true, as this is what common sense dictates.

(88) a. If the canals are *not* frozen, the city does *not* send out icebreaker boats.

b. If it does *not* snow, the city does *not* send out snowplows.

Yet if this is what we can commonly expect, the conditional in (87a) gets a CP-reading: roughly, the city sends out icebreakers if and only if the canals freeze. But the conditional in (87b) does not get a CP-reading.

So this is one simple example demonstrating the general point I would like to make. I do not believe that the fact that (87a) but not (87b) gets a CP-reading is due to general and genuine pragmatic reasoning, such as an I-implicture or a scalar implicature. An implicature-based approach would have to explain why it is that in the same context of utterance one conditional

is perfected while the other is not. I am not saying that such an explanation is inconceivable. I am rather saying that an explanation that appeals to intuitions about common normality expectations does explain the difference (as I find: naturally) and does not require lexical strengthening or scales.

Normality Assumptions. Consequently, I propose that at least some if not most cases of CP should be accounted for in terms of *shared normality assumptions*. To be precise, my suggestion is this: in a stereotypical context of utterance of the conditional in (75a) we come to understand that (75b) is true because of what we take to be *normal courses of events*, in essence, because we take (89) to be true.

(75a) If John leans out of that window any further, he'll fall. $A > C$

(75b) If John doesn't, he will not fall. $\overline{A} > \overline{C}$

(89) John will *normally* not just fall out of the window. "normally $\overline{C}$"

More concretely, I believe that a standard context of utterance for the predictive conditional (75a) *will* feature or readily accommodate (i) a shared presupposition (89) that *normally* John will not fall out of the window in the absence of unexpected intervening events, and (ii) a presupposition that *normally* —as if by definition— unexpected events will not occur. Together this will imply the truth of (75b), as it were, as a natural background assumption about the way the actual world is.

It needs to be stressed that if I say that in a normal context of utterance for (75a) the truth of (89) will already be 'presupposed', I mean that interlocutors can safely rely on (89) as shared implicit background information that forms the basis for interpretation. This is then *not*, of course, a presupposition of the utterance (75a) in the standard linguistic sense: it is an assumption about what is a normal causal development, that informs the interpretation of (75a); but it is not that (75a) would only be true or felicitous if this assumption was in place and that it would therefore trigger accommodation in a context where it was not mutually shared understanding that the world normally behaves in such ways.

Deriving Perfection. It then remains to be shown that, as I claim, (75a) and (89) together imply the CP-reading (75b). In general I need to show that $A > C$ and "normally $\overline{C}$", if given a feasible semantics, imply $\overline{A} > \overline{C}$. For the predictive conditional in (75a), we may assume a modal structure $\langle R_w, \preceq_w \rangle$

where R_w contains all the possible ways the actual world w might develop in the near, relevant future —thus abstracting away from temporal matters and the like— and where $\preceq_w$ represents an understanding of normal courses of events, in the sense that the actual world w is expected to develop into a world in the set Min_w if no unexpected events occur. The assumption that "normally $\overline{C}$" then is spelled out as $\Box\overline{C}$ which is true iff

$$\text{Min}_w \subseteq [\![\overline{C}]\!] .$$

Under these semantics, if $\Box\overline{C}$ is true, then there are no worlds in Min_w that make C true. But if $A > C$ is true as well, then all worlds in Min_w must make $\overline{A}$ true. But that means that $\overline{A} > \overline{C}$ is true. In this sense, a shared background assumption that "normally $\overline{C}$" explains conditional perfection, given a suitable semantics of conditionals and normality assumptions.[9]

In sum, I suggest that appeal to shared normality assumptions is the correct explanation for a great number of cases. This is most plausible for predictive conditionals, especially where they express or relate to a causal relation between events, and commissive conditionals with their strong appeal to a commonsense logic of social contract-making. This is also supported by empirical research investigating how readily subjects infer CP-readings: the studies of Newstead et al. (1997) show that where a "natural causal connection" exists between propositions A and C, CP-readings are readily attested; the same holds of conditional promises and threats.

Perfection from Topicality

But even if a 'normality-based' account of CP-readings *is* correct for a vast number of cases, this cannot be the end of the story. In fact, as far as pragmatic theory is concerned, the most interesting observations are not yet covered by appeal to shared normality assumptions. The point is that there are cases of CP that cannot be readily explained in this way. Take for instance the question-answer pair in (90).

(90) a. Bogart: Will you marry me? ?C

b. Lillie: If I have to. $A > C$

c. $\rightsquigarrow$ If I don't have to, I won't marry you. $\overline{A} > \overline{C}$

d. $\rightsquigarrow$ If I don't have to, I might not marry you. $\overline{A} \diamond\!\!\Leftrightarrow \overline{C}$

9. Notice that I did not make *any* further assumptions about the properties of $\preceq_w$.

It appears to be a piece of fairly robust linguistic intuition that in the context of the question (90a), a conditional answer such as (90b) gets a perfected reading as in (90c) or as in (90d) (see Groenendijk and Stokhof 1984). Still, there is no reason why a general, unbiased context should feature a standing normality assumption that Lillie will or will not marry Bogart unless something extraordinary intervenes. (Who are Lillie and Bogart anyway?) Of course, we could stipulate that the necessary normality assumption will be accommodated, but this line of explanation seems defeatist. It certainly does not account for the generalization that nearly all conditionals get a perfected reading if taken as an answer to a (possibly implicit) question after the truth of their consequents.[10]

Appeal to accommodation of a normality assumption also does not directly explain the rest of the general pattern that can be observed, namely that whether a conditional $A > C$ gets a perfected reading directly depends on the question under discussion. Compare the cases (91)–(93).

(91) a. Q: Is Cathy coming to the party? $?C$
b. A: If Aron is. $A > C$

(92) a. Q: Is Cathy coming to the party if Aron is? $?(A > C)$
b. A: Yes. If Aron is coming, Cathy is coming too. $A > C$

(93) a. Q: Is Aron coming to the party? $?A$
b. A: If Aron is coming, Cathy is coming too. $A > C$

Intuitively, a CP-reading of $A > C$ arises in the context of (91), but not in the context of questions (92) and (93) (see also von Fintel 2001*a*). This observation is not readily explained by means of any of the approaches considered so far, be it (i) lexical strengthening, (ii) scalar implicature or (iii) normality expectations. In conclusion, there appear to be cases where topic requirements force CP-readings while appeal to normality assumptions seems implausible. It is in particular this regular pattern of contextual enrichment in the light of a

10. Exceptions to this rule seem to exist:

(1) a. John: Do you want to order pizza again tonight?
b. Mary: If I may decide what to eat.

This example has a character similar to biscuit conditionals (see section 5.3): it is at least not inferred from the answer that Mary does not want to order pizza again tonight; rather she might want to do so irrespective of whether she may decide what to eat and what not. However, the generalization seems to hold, as far as I can see, for epistemic, predictive and commissive conditionals.

question under discussion that is interesting for a general theory of pragmatic interpretation. I would therefore like to enlarge on this issue in the following.

Perfection from Exhaustivity

As CP-readings vary with the topical question under discussion, an explanation in terms of *exhaustive interpretation of answers* suggests itself. A first very brief and informal statement of the idea to approach CP in this way was given by de Cornulier (1983), but it was Groenendijk and Stokhof (1984) who systematically spelled out exhaustive interpretation of answers in connection with their theory of questions. Groenendijk and Stokhof indeed also applied their theory to conditionals in order to explain CP (on a par with exclusive readings of disjunction, see Groenendijk and Stokhof 1984, pp. 320–328).[11] Exhaustification approaches to implicatures are basically similar in spirit to Q-implicature accounts, but differ from the latter in that they do not refer directly to alternative forms. Rather exhaustification emphasizes the role of the *question under discussion* (cf. van Kuppevelt 1996): for instance, in the light of a question (94a) the answer (94b) is interpreted exhaustively to convey (94c).

(94) a. Who (of John and Mary) came to the party?
b. John did.
c. Mary didn't.

Exhaustive interpretation can be applied to constituent answers as well as to sentential answers. If applied to the latter, it gives a straightforward account for the CP-reading of a conditional as an answer to the question after the truth of its consequent as in (91). To see how this all works, let's briefly review the basic principles of exhaustive interpretation à la Groenendijk and Stokhof.

The Exhaustification Operator. The main idea of the exhaustification approach is to minimize the extension of the question predicate given the truth of the answer. In general then, let's consider a first-order logical language with a finite set of predicate symbols $\mathcal{P}$ of different arities: zero-ary predicate symbols as proposition letters, unary predicate symbols as variables for

11. For overview: von Fintel (2001*a*) reconsiders Groenendijk and Stokhof's account of conditional perfection in the context of more data, and Schulz and van Rooij (2006) spell out exhaustive interpretation in formal detail and apply it to a number of linguistic examples, but do not, as far as CP is concerned, modify or improve on Groenendijk and Stokhof's original analysis.

properties of individuals and so on. Let W be the set of possible worlds, i.e., valuation functions for this language on a given domain $\mathcal{D}$. If $w \in W$ is a possible world, let $w(\varphi)$ be the extension assigned to the formula φ. There are two cases we need to distinguish: (i) if φ contains $n > 0$ free variables, then the extension $w(\varphi) \subseteq \mathcal{D}^n$ is an n-placed relation between individuals from the domain; (ii) if φ is a closed formula without free variables, then $w(\varphi)$ maps onto a truth value, true or false.

Exhaustive interpretation is interpretation that minimizes the extension of a *question predicate*: the question predicate is just a formula of our first-order language, which may either be open, in which case it models a *wh*-question, or closed, so as to model a polar question. I will write T, mnemonic for "topic", as a stand-in for an arbitrary question predicate. A pair of possible worlds $v, w \in W$ is called T-COMPARABLE, $v \cong_T w$, if $v(P) = w(P)$ for all $P \in \mathcal{P} \setminus \{T\}$. A world v is called T-SMALLER OR EQUAL than w, $v \leq_T w$, if $v \cong_T w$ and $v(T) \subseteq w(T)$. If the closed formula A with the semantic value $[\![A]\!] \subseteq W$ is an answer to question predicate T, then the EXHAUSTIVE INTERPRETATION OF A GIVEN T is defined as:[12]

$$exh(A, T) = \{w \in [\![A]\!] \mid \neg\exists v \in [\![A]\!] \; v <_T w\}\,. \tag{5.7}$$

EXAMPLE. Consider a basic treatment of the question-answer pair in (94): we would assume that the question predicate T is the formula

$$\text{Come}(x) \wedge (x = \text{John} \vee x = \text{Mary})$$

with one free variable. We then basically only need to distinguish four types of possible worlds according to whether individuals John and Mary are in the extension of the predicate $\text{Come}(x)$. These four types of worlds are then ordered according to the extension they assign to T, as shown in figure 5.1. In this figure, an arrow from one type of world to another indicates a smaller extension of the question predicate. The semantic meaning of the answer in (94b) is indicated by a shaded area to include only worlds of type w_2 and w_4. Consequently, the T-smallest worlds where (94b) is true are worlds of type w_2, indicated in the figure by a thicker black circle. This is the intuitively correct prediction (that only John but not Mary came to the party), and it illustrates the basic workings of the exhaustivity operator.

12. This is not the original formulation of Groenendijk and Stokhof's exhaustifivity operator, but the reformulation in terms of minimal models given by Schulz and van Rooij (2006). There are minor technical differences that do not play a role here.

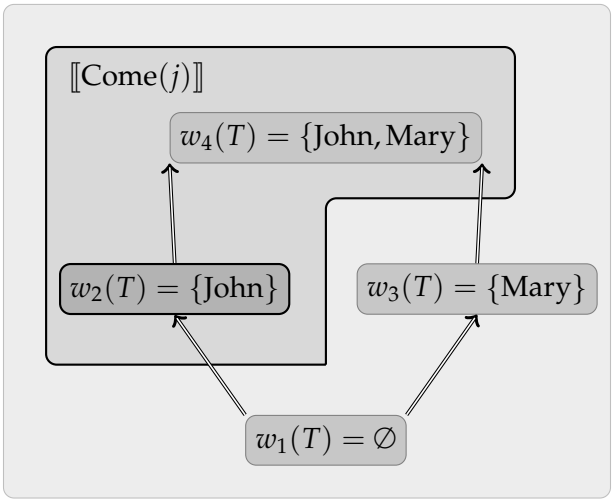


Figure 5.1: Minimal worlds for exhaustive interpretation of example (94)

Perfection from Exhaustivity. Let us turn to Groenendijk and Stokhof's exhaustification-based account of CP. We would like to derive a CP-reading of $A > C$ for question predicate $?C$ and we would like to derive no CP-reading under question predicates $?(A > C)$ and $?A$.[13] Since these are all polar questions, the first thing to do is to specify the order $<_T$ for closed formulas T (which we actually have not yet done). The order $<_T$ should compare worlds with respect to the extension of T. The extension of a closed formula is a truth value. The question is then when is a world T-smaller than another if we compare possible truth values of T? Groenendijk and Stokhof assume that the extension of a closed formula is given as:

$$w(T) = \begin{cases} \{\emptyset\} & \text{if } T \text{ is true in } w \\ \emptyset & \text{if } T \text{ is false in } w \end{cases} \tag{5.8}$$

and consequently receive that $v <_T w$ just in case T is true in w and false in v.

With this assumption in place we can check the predictions in individual cases. Let's start with the case in (91) where a conditional $A > C$ is an answer to question predicate C. Since Groenendijk and Stokhof assume a material implication analysis of conditionals, we need to distinguish four types of worlds, as in figure 5.2, according to the truth value assigned to A and C. These worlds are then ordered with respect to the extension of C as indicated

13. Here, we can save notation and assume that both A and C are zero-ary predicate symbols, i.e., proposition letters.

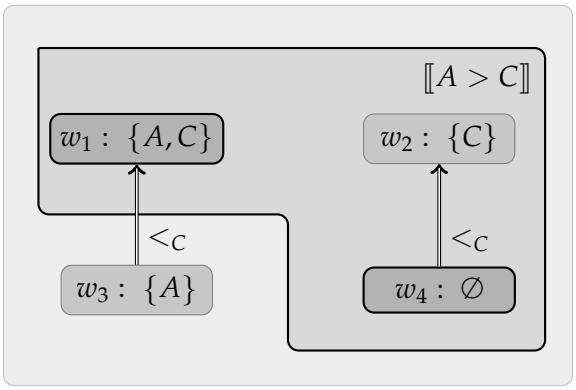


Figure 5.2: C-minimal worlds for exhaustive interpretation of $A > C$, as in example (91)

by the arrows. In line with assumption (5.8), we get $v <_C w$ just in case C is true in w and false in v. The figure indicates the type of all those worlds in which conditional $A > C$ is true under a material implication analysis: $\{w_1, w_2, w_4\}$. The graphic makes it easy to see that exhaustive interpretation of $A > C$ thus excludes worlds of type w_2 and leaves only w_1 and w_4 as the exhaustive interpretation of the conditional. Groenendijk and Stokhof's account effectively derives a material biconditional reading.

A similarly satisfactory result is obtained for the case in (92). As Groenendijk and Stokhof show, if $A > C$ is an answer to the question $?(A > C)$, the conditional will not be exhaustified and maintains its material implication meaning. This is a direct consequence of the general result that answers "yes" and "no" will not receive an exhaustive reading (see Groenendijk and Stokhof 1984, pp. 322-323).

So, thus far predictions are good. But unfortunately the predictions of the exhaustification operator for the case (93), where $A > C$ appears as an answer to the question $?A$, are incorrect. Following the same logic as before, we would minimize the truth-value of the question predicate A in the interpretation of the conditional $A > C$. The operation is graphically depicted in figure 5.3 and shows that the approach predicts that the conditional is equivalent to a straight "no" answer in this case.

Error Analysis. Taking a small step back and reflecting on these predictions, it seems fair to say that the stipulation in (5.8) is actually doing most of the work in the derivation of CP under topic $?C$: the definition in (5.8) aligns the exhaustification operator in (5.7), at least formally speaking, with

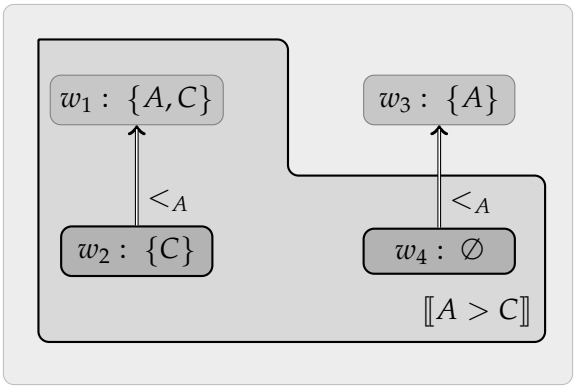


Figure 5.3: A-minimal worlds for exhaustive interpretation of $A > C$, as in example (93)

interpretation in 'normal worlds' where C is assumed not to hold, as given in section 5.2.2.

Seen in this light, we might actually start to doubt whether it is conceptually adequate to base the order $<_T$ for closed formulas T on the stipulated extensions in (5.8). Admittedly, there is an intuitive rationale behind ordering worlds in terms of extensions if the question predicate T is an *open* formula. But if T is a closed formula things are not that intuitive: in what sense is a world that makes T false more extension-minimal than one where T is true?

I suspect that it is the somewhat stipulative character of an ordering based on (5.8) which is responsible for the erroneous prediction of case (93). Put in slightly provocative terms, perhaps exhaustive interpretation successfully derives CP-readings under topic $?C$ only accidentally, as a conceptually sound derivation under topic $?C$ would carry over flawlessly to the topic $?A$.

OUTLOOK. The next two sections offer an alternative account that aims to overcome exactly these problems. I will first spell out the proposal in generally accessible terms in section 5.2.3. This section basically appeals to commonsense intuitions and gives a plausibility account for the derivation, respectively non-derivation, of CP-readings. The following section 5.2.4 backs up the plausibility account with a concrete game theoretic model. This is however conceptually fairly involved and so some readers may happily content themselves with reading only section 5.2.3.

5.2.3 Rationalizing Indirectness

The main idea which is to be spelled out here and refined in the following section is very simple. Take a polar question ?T and think of T as a stand-in for either antecedent A or consequent C of a conditional. In the light of a question ?T, I will treat the conditional $A > C$ as an *indirect answer* which has to be *rationalized against a direct answer*: basically, the hearer asks himself why the speaker has not just said "yes" or "no" when he hears $A > C$ as an answer to ?T.[14] Similar to scalar reasoning, the hearer then rules out all those worlds from the semantic interpretation of $A > C$ where a direct answer is true. If $[\![A > C]\!]$ be the set of worlds where the conditional $A > C$ is (non-trivially) true, then this idea spells out roughly as:

$$\text{Intpr}(A > C, T) = \{w \in [\![A > C]\!] \mid \text{"yes" and "no" are not true in } w\} \quad (5.9)$$

But, of course, the direct answers "yes" and "no" are different for different contextual questions ?$T \in \{?A, ?C\}$. Effectively, different topical questions induce different *contextual alternatives* with which to compare $A > C$. If we spell this out more carefully, we find that in the one case we derive CP-readings, while in the other we don't. This is the whole idea in a nutshell.

Direct Answers as Modal Statements

If we want to make the operation in (5.9) precise, we need to pin down how to analyze the answers "yes" and "no" for comparison with the conditional. It turns out that a too naïve approach soon runs into a formal impasse. The problem is obvious. If T is a proposition that is either true or false, and if the direct answer "yes" ("no") means that T is true (false), then this reasoning eliminates *all* worlds from $[\![A > C]\!]$:

$$\begin{aligned} \text{Intpr}'(A > C, T) &= \{w \in [\![A > C]\!] \mid T \text{ and } \overline{T} \text{ are not true in } w\} \\ &= \emptyset \end{aligned}$$

Instead, I suggest, direct answers "yes" and "no" should be interpreted also *in the light of the conditional* and should thus be susceptible to all modal distinctions the conditional introduces. That means that, roughly put, to interpret

14. It is not crucial to agree with my choice of words calling "yes" and "no" the (only) 'direct answers' to a polar question. All that matters for the present concern is that a polar question (usually, normally) makes a simple "yes" and a simple "no" much more salient answers than a conditional $A > C$. This is the crucial intuition, not the terms 'direct' and 'indirect.'

$A > C$ in the light of $?T$ is to rule out worlds from the semantic interpretation of $A > C$ where the *modalized* expressions $\Box T$ and $\Box\overline{T}$ are true:

$$\text{Intpr}^{\Box}(A > C, T) = \{w \in [\![A > C]\!] \mid \Box T \text{ and } \Box\overline{T} \text{ are not true in } w\} \quad (5.10)$$

It is exactly this move from contextual alternative T to $\Box T$, so to speak, that the following section will back up with an involved argument based on the dynamics of awareness. The point can, however, also be made in intuitive terms. Here is my argument based on epistemic conditionals.

If a question like (95a) is answered with a straightforward "yes" or "no" we usually do not take this to be information about the speaker's epistemic state, but rather about the actual world.

(95) a. Q: Did Cathy come to the party?
b. A: Yes.
c. A: No.
d. A: If Aron did.

But if we interpret the conditional answer in (95d), accommodating the epistemic dimension seems unavoidable. And once the interpreter is sensitized to these modal distinctions due to the conditional, looking back at answers "yes" and "no" from this point of view also means to interpret answers (95b) and (95c) as saying that the speaker *knows* that T is true, respectively that the speaker *knows* that T is false. Hence, although on their own (95b) and (95c) would be taken to refer directly to matters of truth and falsity of the actual world, if evaluated against an epistemic background setting —which the conditional introduces— the direct answers should also be interpreted in relation to the speaker's epistemic state.

A similar argument applies to other kinds of conditionals which may introduce other kinds of modality. Generally speaking, if we assume that conditionals are evaluated on a modal structure $\langle R_w, \preceq_w \rangle$, then the direct answers "yes" and "no" to topical question $?T$ should also be evaluated with respect to this modal structure as $\Box T$ and $\Box\overline{T}$ respectively. We would then rule out all those worlds from $[\![A > C]\!]$ where $\Box T$ and $\Box\overline{T}$ are true, instead of those where T and $\overline{T}$ are true.[15]

15. This works fine unless we explicitly restrict the class of modal structures $\langle R_w, \preceq_w \rangle$ to include only orderings $\preceq_w$ for which there is exactly one $\preceq_w$-minimal world. So, in particular, we would run into the same problem as before if we assumed that the ordering had to satisfy *strong centering*. However, this is not a problem as long as we deal with indicative conditionals for which such a restriction does not seem desirable to begin with.

Deriving Perfection

It remains to be shown that the interpretation operation sketched in (5.10) really derives (weak) CP-readings under topic C and not under topic A, as we would like it to. I will content myself with presenting only abstract, formal results and invite the reader to check these against her favorite application of the assumed conditional semantics.

TOPIC ?C. Take a topical question ?C. We would like to show that the attested weak CP-reading (that it is not the case that $\overline{A} \boxRightarrow C$) must hold whenever $A \boxRightarrow C$ is true, and $\Box C$ and $\Box\overline{C}$ are false. So suppose that in a world w with modal structure $\langle R_w, \preceq_w \rangle$ the conditional $A \boxRightarrow C$ is true, and that it is also the case that $\overline{A} \boxRightarrow C$ is true. This implies that all worlds in Min_w actually make C true, which contradicts the assumption that $\Box C$ is false. Hence, if $A \boxRightarrow C$ is true and it is not the case that $\Box C$ is true, then the weak CP-reading is derived as desired. It is moreover plain to see that the strong CP-reading in (79) is not ruled out by this reasoning. This derivation furthermore does not require non-triviality or any other special properties of the modal structure.

TOPIC ?A. For an argument why the same pragmatic reasoning is not strong enough to generally derive a CP-reading under ?A first notice that ruling out worlds where $\Box A$ and $\Box\overline{A}$ are false is equivalent to a *strong* non-triviality presupposition that $\Diamond A$ and $\Diamond\overline{A}$. But a world w with $\text{Min}_w \cap A \neq \emptyset$ and $\text{Min}_w \cap \overline{A} \neq \emptyset$ may also have $\text{Min}_w \subseteq C$, thus making $A > C$, as well as $\overline{A} > C$ true. Consequently, the topic-dependent pragmatic strengthening will not yield CP-readings unless these are forced by something else, such as world knowledge or other contextual assumptions.

FROM WEAK TO STRONG PERFECTION. So (5.10) properly derives weak CP-readings under topic C, but not under topic A. Still, often strong CP-readings spring more readily to mind than weak CP-readings if $A > C$ is an answer to ?C as in (90). I therefore suggest to think of the contrast between strong and weak CP-readings in parallel to expert and inexpert epistemic readings of scalar implicatures (see section 3.2). Here is my argument.

Suppose that the modal structure $\langle R_w, \preceq_w \rangle$ captures plain epistemic modality, Hintikka-Kripke-style, such that R_w is a set of doxastic alternatives and $\preceq_w$ is just the total relation on R_w. With our non-triviality presupposition ($R_w \cap A \neq \emptyset$) in place, a conditional is non-trivially true in four different

kinds of such epistemic states, which we can represent using notation from lifted signaling games as follows (see section 3.2):

$$[\![A > C]\!] = \left\{ t_{[\mathsf{AC}]},\ t_{[\emptyset,\mathsf{AC}]},\ t_{[\mathsf{C},\mathsf{AC}]},\ t_{[\emptyset,\mathsf{C},\mathsf{AC}]} \right\}. \tag{5.11}$$

A state $t_{[\mathsf{C},\mathsf{AC}]}$, for instance, is a state in which the speaker thinks it is possible that either only C is true or that A and C are both true. The interpretation operator in (5.10) rules out two of these states and leaves us with:

$$\begin{aligned} \text{Intpr}^{\Box}(A > C, C) &= \left\{ w \in [\![A > C]\!] \mid \Box C \text{ and } \Box\overline{C} \text{ are not true in } w \right\} \\ &= \left\{ t_{[\emptyset,\mathsf{AC}]},\ t_{[\emptyset,\mathsf{C},\mathsf{AC}]} \right\} \end{aligned}$$

and this amounts to a weak CP-reading because the state $t_{[\emptyset,\mathsf{C},\mathsf{AC}]}$ is included here. However, adopting our previous notion of speaker expertise, we say that the speaker is more of an expert in state $t_{[\emptyset,\mathsf{AC}]}$ than in $t_{[\emptyset,\mathsf{C},\mathsf{AC}]}$ because she entertains strictly fewer possibilities in the former than in the latter. It now seems defensible to assume that interpretation favors more 'minimal states' in this sense, as an assumption about speaker expertise or 'simpler models' in general. Hence, strong CP-readings spring more readily to mind.

Summary. To take stock, in this section I have argued that if we translate an (implicitly) assumed contextual question into an expectation of a direct answer, then we can use scalar-like reasoning to rule out certain parts of the meaning of a conditional, namely those states for which a direct answer would have been true. This accounts for the context-dependence of CP-readings on an intuitive basis, but the question remains what kind of account this scalar-like reasoning process actually is. The following section addresses this concern by giving a game theoretic rationale for the suggested reasoning process.

5.2.4 Forward Induction under Awareness Dynamics

I suggest that the kind of context-induced scalar reasoning I have spelled out in the last section is corroborated by the general principles of model construction *ex post* that were given in section 3.1.2, if we take into account different conceptualizations of the context of utterance the interpreter has after hearing different messages. This latter aspect is modelled by ascribing different states of *awareness* to the receiver and incorporating such awareness dynamics into pragmatic reasoning.

	$\Pr(t)$	$a_{\mathtt{T}}$	$a_{\overline{\mathtt{T}}}$	$m_{\mathtt{T}}$	$m_{\overline{\mathtt{T}}}$
$t_{\mathtt{T}}$	$1/2$	1,1	0,0	$\surd$	$-$
$t_{\overline{\mathtt{T}}}$	$1/2$	0,0	1,1	$-$	$\surd$

Figure 5.4: $G_{?T}$ – aka 'game whether T'

The Game whether T. We should start with a general question: what does a signaling game look like that models a contextual question under discussion? As I have argued in section 3.1, signaling games actually model a question under discussion in the set of available interpretation actions and the structure of the payoffs. The most straightforward implementation of a contextual question $?T$, is the game $G_{?T}$ —read as 'game whether T'— in figure 5.4. There are two states $t_{\mathtt{T}}$ and $t_{\overline{\mathtt{T}}}$ which capture all those distinctions that are relevant given that interlocutors are interested in whether the proposition T is true. By the same token, the set of messages in this game is the set $M = \{m_{\mathtt{T}}, m_{\overline{\mathtt{T}}}\}$, basically saying "yes" and "no" to the contextual question $?T$.

Games with Evolving Awareness. Suppose that this is the game that is played under a contextual question $?T$, and remember that we take $T \in \{A, C\}$ as a placeholder for either the antecedent or the consequent of a conditional whose interpretation we are interested in. Then, obviously, the conditional $A > C$, which we would like to have interpreted in the light of question $?T$, is not even in the set of available messages of the game $G_{?T}$. This is how it should be, I propose, in line with the intuition which I have argued for in the previous section already, namely that $A > C$ seems, in a sense, unexpected, or at least less expected as an answer to polar $?T$ than the answers "yes" and "no."

The game model that I will endorse here to capture this situation is a dynamic game with unawareness (Feinberg 2004, 2005; Heifetz et al. 2009). I will assume that the receiver is initially unaware of the message $m_{\mathtt{A>C}}$ and remains unaware of it if he observes a direct answer $m_{\mathtt{T}}$ or $m_{\overline{\mathtt{T}}}$. In that latter case he will believe that the game to be played is $G_{?T}$ and that's that. But the observation of message $m_{\mathtt{A>C}}$ —or any other unexpected move by the sender— will make the receiver aware of this message and will make him accommodate his representation of the game, i.e., his conceptualization of the context of utterance. I will assume that if the receiver observes message $m_{\mathtt{A>C}}$ he will come to believe that the signaling game is no longer the game

$G_{?T}$, but rather a game $G^{+}_{?T}$ which is derived from $G_{?T}$ in such a way that it additionally includes at least the unexpected message $m_{\mathtt{A}>\mathtt{C}}$ together with any other necessary changes.

One such necessary change concerns the set of states that ought to be distinguished in $G^{+}_{?T}$. In the original game $G_{?T}$ we only had two states $t_{\mathtt{T}}$ and $t_{\overline{\mathtt{T}}}$, but clearly this is not the level of granularity against which the message $m_{\mathtt{A}>\mathtt{B}}$ should be evaluated: the conditional as such introduces additional distinctions in the set of states that have not been taken into account in $G_{?T}$. Let us assume that $G^{+}_{?T}$ contains the alternative messages

$$M^{+} = \{m_{\mathtt{T}}, m_{\overline{\mathtt{T}}}, m_{\mathtt{A}>\mathtt{C}}\}\,.$$

Then our standard procedure for canonical model construction requires to consult four initially possible state distinctions:

	$m_{\mathtt{T}}$	$m_{\overline{\mathtt{T}}}$
t_1	$\surd$	$\surd$
t_2	$\surd$	$-$
t_3	$-$	$\surd$
t_4	$-$	$-$

Recall that then t_2, for example, is the set of all worlds where $m_{\mathtt{A}>\mathtt{C}}$ and $m_{\mathtt{T}}$ are true but where $m_{\overline{\mathtt{T}}}$ is false.

But now the question arises how to interpret the messages $m_{\mathtt{T}}$ and $m_{\overline{\mathtt{T}}}$ in the game $G^{+}_{?T}$. If we take these messages to say that T is true, respectively false, *as such*, then the only consistent states are t_2 and t_3. But this is not what we should do, and it is here that I can further motivate my previous suggestion to interpret messages $m_{\mathtt{T}}$ and $m_{\overline{\mathtt{T}}}$ as modalized statements that are related to the same modal structure that an evaluation of $m_{\mathtt{A}>\mathtt{C}}$ requires. Here is the argument from 'reasoning about foregone unawareness.'

Reasoning with Unawareness. To repeat for clarity, the idea of modelling reasoning about dynamic unawareness is this. Since we assume that the contextual question under discussion is $?T$, the receiver reasons about the game $G_{?T}$ after hearing messages $m_{\mathtt{T}}$ and $m_{\overline{\mathtt{T}}}$ and will not make any of the additional distinctions that *some* unexpected signal may force upon him. Yet, in the game $G^{+}_{?T}$, i.e., from a perspective of broader awareness and more fine-grained distinctions, the receiver *can* reason about his own (counterfactual) state of limited awareness in $G_{?T}$. In general, there is a natural asymmetry

in the reasoning power of agents in games with dynamic awareness: from a state of awareness an agent can reason about his hypothetical beliefs, views and dispositions to act had he been unaware of certain contingencies, but in a state of unawareness an agent cannot —as if by definition— reason about the beliefs he would hold and the actions he would choose in case he had been aware of contingencies that he is in fact not aware of.

In order to implement the reasoning capabilities of agents with different awareness states into the IBR model, I will follow in particular the formalization of awareness dynamics in extensive games developed by Feinberg (2004; 2005). I will assume that each strategic receiver type comes in two versions: either he is aware of the conditional $m_{\mathtt{A}>\mathtt{C}}$ and the game $G^{+}_{?T}$, or he is not, depending on whether he observed the conditional or a direct answer. In effect, we can then apply the IBR model without modification to the trivial game $G_{?T}$ with receiver types R_k as before. In the game $G^{+}_{?T}$, on the other hand, we will have to assume receiver types R^{+}_{k} with extended awareness. Receiver types R^{+}_{k} not only believe that the context of utterance is modeled by $G^{+}_{?T}$ but also know about their foregone state of unawareness and they can conceive of how they would have reasoned and acted had they been of the unaware type R_k.[16]

Awareness Evolution Triggers Forward Induction. The next question to be settled then is how to characterize the reasoning behavior of the aware receiver types R^{+}. Following in particular Heifetz et al. (2009), I argue that we should analyze the receiver as being *surprised* by the message $m_{\mathtt{A}>\mathtt{C}}$: since from unawareness he had in a certain sense expected that a direct answer would be sent, any message that is not a direct answer to the question under discussion is a surprise message that needs to be rationalized *ex post*.[17] But that means that the message $m_{\mathtt{A}>\mathtt{C}}$ should be a surprise message already for the first occurring receiver types, namely R^{+}_{0} and R^{+}_{1}. As a result, unlike in the basic version of the model without dynamic awareness, already these receiver types need to rationalize the use of surprise message $m_{\mathtt{A}>\mathtt{C}}$.

16. There are many more interesting subtleties in properly fitting reasoning about unawareness into the IBR model. It may seem natural, for instance, to rule that an aware receiver type should not reason (much) higher up the IBR sequence for the unaware game. Such complications, however, don't interfere with the relatively simple application here.

17. It is not actually necessary to imagine the receiver to explicitly or implicitly believe that the game is $G_{?T}$ before he observes $A > C$ in order to be 'surprised' in this technical sense. The utterance $A > C$ can be entirely out-of-the-blue and still the receiver can construct the context as about the question $?T$ in which the conditional is a 'surprise.'

	$\Pr(t)$	a_2	a_3	a_4	$m_{\mathtt{T}}$	$m_{\overline{\mathtt{T}}}$	$m_{\mathtt{A}>\mathtt{C}}$
t_2	$1/3$	1,1	0,0	0,0	$\surd$	–	$\surd$
t_3	$1/3$	0,0	1,1	0,0	–	$\surd$	$\surd$
t_4	$1/3$	0,0	0,0	1,1	–	–	$\surd$

Figure 5.5: $G^{+}_{?T}$ – the 'game whether T' after accommodating $A > C$

This implies that already the basic state distinctions in $G^{+}_{?T}$ should be sensitive to the way $m_{\mathtt{T}}$ and $m_{\overline{\mathtt{T}}}$ would have been interpreted by an unaware receiver. An aware receiver type knows that an unaware type who observes $t_{\mathtt{T}}$ ($t_{\overline{\mathtt{T}}}$) comes to believe that the sender knows that T is true (false). Hence it is these meaning differentiations that inform the construction of $G^{+}_{?T}$. The states in this game should therefore to be conceived as follows:

$$\begin{aligned}
t_1 &= \{w \in [\![A > C]\!] \mid \Box T \text{ and } \Box\overline{T} \text{ are true in } w\} \\
t_2 &= \{w \in [\![A > C]\!] \mid \Box T \text{ is true and } \Box\overline{T} \text{ is false in } w\} \\
t_3 &= \{w \in [\![A > C]\!] \mid \Box T \text{ is false and } \Box\overline{T} \text{ is true in } w\} \\
t_4 &= \{w \in [\![A > C]\!] \mid \Box T \text{ and } \Box\overline{T} \text{ are false in } w\}
\end{aligned}$$

Under this interpretation of $m_{\mathtt{T}}$ and $m_{\overline{\mathtt{T}}}$ only the state m_1 is inconsistent. Our signaling game model $G^{+}_{?T}$ is the context model in figure 5.5, from which it is obvious that the IBR model will assign interpretation t_4 to the message $m_{\mathtt{A}>\mathtt{C}}$. We thus derive the exact same prediction as before under the more intuitive scalar-like reasoning outlined in the last section. What the game theoretic model adds to the picture is a justification for exactly this kind of scalar-like reasoning: as an indirect answer to a contextual question under discussion the receiver constructs a context representation after the fact that accommodates his own foregone unawareness, i.e., he integrates into his own aware representation of the context how he would have interpreted messages if he had remained unaware.

SUMMARY. I suggest to conclude positively that the mission's objectives have been met. We wanted to account for the topic dependence of CP-readings and we have done so first in intuitive terms and then backed up by a rather involved game theoretic model that implemented a notion of indirectness of answers by awareness dynamics of the receiver. In effect, the model thus mimicked scalar reasoning with a contextual scale in which the conditional

is compared with the direct answers to the contextual question under discussion. This, however, is only *as-if*-scalar reasoning, because the game model does not rely on a fixed scale but rather accounts for the contextual adoption of a set of alternative expressions by awareness dynamics, as proposed in several recent accounts in the rational choice literature.

5.3 Unconditional Readings

Conditional perfection, the topic of the last section, is in a certain sense the mirror imagine of another interesting phenomenon in the interpretation and use of conditionals: some conditionals not only do not get a CP-reading, but even receive what I would like to call *unconditional readings*. A particularly representative instance of such conditionals is the class of *biscuit conditionals* (BCs) — so-called after Austin's example (65a), repeated here.

(65a) There are biscuits on the sideboard if you want them.

This conditional is remarkable because it relates propositions "there are biscuits on the sideboard" and "you want some biscuits" in a conditional construction, although these are by common sense conditionally unrelated, as far as their content is concerned: whether there are biscuits on the sideboard at the present moment is not dependent on whether the addressee would like some or not. It is in this sense that I speak of unconditional readings of conditionals, and its such unconditional readings that I would like to deal with in this section.

The main idea which I would like to put forward here is that unconditional readings can be derived from a standard semantics of conditionals together with a contextual assumption of *conditional independence* of propositions. I will give a suitable formal notion of conditional independence and relate it to existing notions, such as logical and probabilistic independence. Not all unconditional readings deserve or require an account of this kind, though. To delineate which ones do and which ones do not is therefore the secondary objective of this section.

This section is then decidedly *not* exclusively about biscuit conditionals. Philosophers and linguists alike have frequently adopted the view that BCs are a special subspecies of conditionals that can be singled out by peculiar syntactic and perhaps intonational properties. I will elaborate on some of the properties of BCs in section 5.3.1 and review some influential and recent accounts of BCs which also aim to derive unconditional readings. Still, as

I would like to show in section 5.3.2, unconditional readings do not occur only for BCs with their distinct syntactic properties, but also for seemingly standard conditionals. This, I argue, casts doubt on the relevance of accounts of unconditional readings that are based on special properties of BCs. I will then offer a very general pragmatic explanation for unconditional readings in section 5.3.3 and finish in section 5.3.4 with a game theoretic explanation of the discourse effects of BCs and related constructions.

5.3.1 Biscuit Conditionals

Biscuit conditionals are conditionals named after the example in (65a) which have been discussed as special cases of conditionals from a variety of angles under a variety of names.[18] Further examples are the sentences in (96).[19]

(96) a. If I may say so, this is boring.

b. Her dress is too German for my taste, if you know what I mean.

c. If we now turn to the last agenda item, fund cuts are tremendous.

There are striking intuitive differences between these examples and more standard conditionals. In a rough first approximation, the intuitive difference seems to be that (i) BCs appear to somehow convey the unconditional truth of their consequents and (ii) the antecedents of BCs relate in some fashion to matters of felicity or relevance of the consequent material.

Characterizing Biscuits. How exactly to delineate these intuitive differences is, however, a rather delicate matter, and it is here that we very clearly see mere description of the data blend into theorizing. Some authors have claimed that the antecedent material gives conditions on the very speech act performed by the consequent:

> "[P]ragmatic *if* is a typical conditional for speech acts: it specifies the conditions —of a context unknown to the speaker— under which a speech

18. Here are some of the labels used by various authors, often indicative of the respective author's preferred analysis: non-conditional conditionals (Geis and Lycan 1993), speech-act conditionals (van der Auwera 1986; Sweetser 1990), relevance conditionals (Iatridou 1991), metarepresentational conditionals (Noh 1998) or non-interference conditionals (Bennett 2003).

19. We could be more thorough and further distinguish subtypes in this vaguely defined set. Günthner (1999), for instance, differentiates meta-communicative conditionals like (96a) and (96b) and discourse-structuring conditionals like (96c) from relevance conditionals like (65a).

> act should count. That is, I inform you of the following fact which would be of use to you in the event that you need me."
>
> (van Dijk 1979, p. 455)

> "[T]he preferred reading [of a BC] has the adverbial modifying the act of stating, informing, etc." (Davison 1983, p. 505)

Others have claimed instead that the antecedent material gives conditions under which the consequent material —be that the speech act associated with the consequent or its semantic content— is relevant in some appropriate sense:

> "Nevertheless, although in a non-integrative conditional [i.e., in a BC] the truth of the protasis is not sufficient for the truth of the apodosis, the truth of the protasis is a sufficient condition for the relevance of the speech act vehicled through the apodosis." (Köpcke and Panther 1989, p. 694)

> "[T]he *if*-clauses in [BCs] specify the circumstances in which the consequent is relevant (in a vague sense, also subsuming circumstances of social appropriateness), not the circumstances in which it is true."
>
> (Iatridou 1991, p. 51)

Conditional Speech-Acts. Both assessments have motivated analyses of BCs as some sort of *conditional speech-acts*. In crude outline, a generic instance of this explanation scheme would either, as in (97a), postulate an elliptical performative (cf. Rutherford 1970; van der Auwera 1986; Iatridou 1991) or, as in (97b), some abstract illocutionary force operator (cf. Davison 1983; Sweetser 1990; DeRose and Grandy 1999).

(97) a. If you want some, (I hereby say to you that) there are biscuits on the sideboard.

b. If you want some, ASSERT("there are biscuits on the sideboard").

Even where we neglect the intricacies of individual proposals, with their respective merits and flaws, it is still fair to say what is unappealing about any such account. Firstly a conceptual point: conditional speech-acts, if taken seriously, are very peculiar entities —where else in life do you perform your actions conditionally?— whose properties can only be assessed via exactly those sentences' meanings whose meaning they are to explain.[20] Secondly, it

20. I don't want to commit myself to claiming that there are no conditional speech-acts whatsoever, but I certainly believe that it takes very peculiar, stylized circumstances to have a speech act come out as (if it was) conditionally performed.

is implausible to treat a case like (98) —and especially a past tensed one like (99)— as a multiply performed speech-act, as the given paraphrases would suggest (see Siegel 2006, for related criticism).

(98) a. If/Whenever you need anything later, my name is James.
b. If/Whenever you need anything later, (I hereby say to you that) my name is James.
c. If/Whenever you need anything later, ASSERT("my name is James").

(99) a. Ah, living in California was great! If/Whenever we wanted to go for a swim, the sea was just a five minute ride away.
b. If/Whenever we wanted to go for a swim, (I hereby say to you that) the sea was just a five minute ride away.
c. If/Whenever we wanted to go for a swim, ASSERT("the sea was just a five minute ride away").

POTENTIAL LITERAL ACTS. Against naïve conditional speech-act accounts, Siegel (2006) suggests to analyze BCs in terms of what she calls *quantification over potential literal acts*.[21] Eventually, Siegel offers the paraphrase in (100) as her analysis of the BC in (65a).

(100) If you want them, there is a (presupposed relevant, salient, and otherwise felicitous) potential literal assertion with the propositional content "there are biscuits on the sideboard."

Opposed to the rather strong conditional speech-act accounts, this account is fairly weak, both semantically and pragmatically. For one, Siegel's account predicts that BCs are always true semantically —potential literal acts should always exist in abstract space—, and so Siegel has to argue that BCs may or may not have presupposition failures (of varying severity), as there may not always be relevant, salient, and otherwise felicitous potential literal acts (for criticism, see also Predelli 2007). For another, there is still quite a gap to be bridged from the existence of a potential literal act —be that relevant or not— to the actual performance of a concrete speech act. This is a problem

21. Here is how Siegel characterizes potential literal acts: "these semantic objects are *not* literally *acts*, not things that people actually do. They lack the contextual specifics of actual speech acts: a speaker, an addressee, an appropriate context. [...] They are abstract objects consisting only of propositional content and whatever illocutionary force potential can be read directly from their morphosyntactic form, not necessarily the actual illocutionary act that might be performed." (Siegel 2006, p. 170)

because, as Ebert et al. (2008) as well as Scheffler (2008*a*,*b*) argue, examples such as (101) seem to show that the speech act associated with the consequent is always performed unconditionally on the truth of the antecedent.[22]

(101) a. If I don't see you anymore, I hope you enjoy your holiday!

b. If you don't want to watch the movie, the gardener is the killer.
(Ebert et al. 2008, (3))

c. If the congregation is ready, I hereby declare you man and wife.
(Ebert et al. 2008, (4))

Unconditional Speech-Act "*C*". In particular, Scheffler (2008*a*,*b*) suggests that a BC $A > C$ is to be analyzed as (i) asserting C outright and (ii) *conventionally implicating* that "if A, then UTTER(C)." This analysis is motivated by data showing that BCs behave similar to certain other attested conventional implicature items with respect to embeddability under, for instance, negations, question operators and attitude verbs.

Ebert et al. (2008), on the other hand, advance a speech-act conjunction analysis of BCs according to which an utterance of the BC $A > C$ performs (i) an act of referring to a possible world with the antecedent A, and (ii) the speech act associated with the consequent C. This proposal is corroborated with data showing how BCs behave similar to certain topic constructions with respect to binding of pronouns in the consequent by quantifiers in the antecedent.

Taken together, both of these accounts derive unconditional readings of BCs by assuming that the consequent is asserted (or that another veridical speech-act with the content that C is performed). To support this explanation, both accounts seek to work out special characteristics of BCs, such as embeddability (Scheffler) or binding properties (Ebert et al.).

Does this suffice as a satisfying account of unconditional readings as such? I believe the answer is "no." I believe that unconditional readings arise also independently of special semantic or syntactic properties of certain conditionals. Unconditional readings arise by pragmatic strengthening whenever propositions occur in a conditional construction that are not conditionally related by commonsense, be that causally, evidentially, logically or in any other conceivable way. In order to support this claim, the next section will review data that shows that the question whether a conditional has an unconditional

22. A nicely twisted way of framing this argument would be to say that you cannot avoid insulting somebody by hedging "If this does not offend you, you're a total idiot!"

reading is orthogonal to the question whether the conditional satisfies basic properties associated with BCs.

5.3.2 Unconditionality Beyond Biscuits

The mistaken idea that BCs are the only conditionals that receive unconditional readings readily suggests itself, especially in the light of the widespread but dubious conviction that BCs form a syntactically neatly delineated subclass of conditionals. A key argument in favor of this latter hypothesis revolves around the observation that in certain languages, such as Dutch or German, an English conditional like (102), which is ambiguous between a conditional and an unconditional reading, is disambiguated in Dutch and German by word order of the consequent. While both sentences in (103) and (104) translate into (102), the variants with main clause verb-second (V2) word order in (103a) and (104a) get an unconditional reading only; the verb-first (V1) word order in (103b) and (104b), on the other hand, gets a conditional reading only.

(102) If you need me, I'll stay at home all day.

(103) a. Als je me nodig hebt, ik blijf de hele dag thuis.
If you me need, I stay the whole day at home.

b. Als je me nodig hebt, blijf ik de hele dag thuis.
If you me need, stay I the whole day at home.

(104) a. Wenn du mich brauchst, ich bleibe den ganzen Tag daheim.
If you me need, I stay the whole day at home.

b. Wenn du mich brauchst, bleibe ich den ganzen Tag daheim.
If you me need, stay I the whole day at home.

Thus conceived, it is natural to hypothesize that there is a clear syntactic demarcation between standard, truly conditional conditionals and conditionals with an unconditional reading, and that this dividing line falls together with the distinction between standard conditionals and BCs.

This idea is, however, not correct in its generality. Already Köpcke and Panther (1989) dismissed the above hypothesis in its strong formulation because, as they argue, there are (i) V2-cases with conditional readings, and (ii) V1-cases with unconditional readings. Köpcke and Panther give the example (105) as an example where even the V2-variant gets a conditional reading.

(105) a. Wenn er das erfährt, gibt es Ärger.
If he that find out result in it trouble.

b. Wenn er das erfährt, das gibt Ärger.
If he that find out that result in trouble.
'If he finds out about this, there will be trouble.'
(Köpcke and Panther 1989, (45))

Strengthening this point, (106) is an example of my own which does not rely on the questionable topical proform *das* in the V2-variant, but still gets a clear conditional reading.

(106) a. Wenn du auch nur in die Nähe meines Autos kommst, spuck ich
If you also only in the vicinity of my car come, spit I
dir in deine Suppe.
you in your soup.

b. Wenn du auch nur in die Nähe meines Autos kommst, ich spuck
If you also only in the vicinity of my car come, I spit
dir in deine Suppe.
you in your soup.
'If you come anywhere close to my car, I'm going to spit in your soup.'

Intuitively, both variants in (106) express that the speaker is going to spit in the hearer's soup if (and only if) he gets near her car.

I agree that one could analyze the sequence in (106b) as two separate speech acts, because individual *if*-clauses can indeed occur on their own, especially to make threats. The subsequent main clause could then be taken as a standalone assertion which is restricted in scope by a general mechanism of modal subordination (Roberts 1989). I would even endorse such an analysis because it drives the mills of my argument. The most natural explanation for why we restrict the second assertion by modal subordination (if that is what we are doing), is because that makes sense *pragmatically*: that the spitting can be prevented by staying away from the speaker's car is an absolutely natural idea in a normal context of utterance of (106). But that means that whether taken as a single conditional or not, common sense takes the involved propositions to be very much conditionally related in this case, so as to even establish a conditional reading *despite* a main clause V2 word order.

Still, it is even more important to my overall concern that there are also conditionals with integrative V1 word order that nonetheless get an unconditional reading. Köpcke and Panther give the examples in (107) and (108), all variants of which were found acceptable by subjects in Köpcke and Panther's survey and all variants of which convey that the consequent holds independently of whether the antecedent does.

(107) a. Wenn Sie mich fragen, es schneit bald.
If you me ask, it snow soon.

b. Wenn Sie mich fragen, schneit es bald.
If you me ask, snow is soon.

c. Wenn Sie mich fragen, dann es schneit bald.
If you me ask, then it snow soon.

'If you ask me, it'll snow soon.' (Köpcke and Panther 1989, (48))

(108) a. Wenn du meine Meinung hören willst, die Aktien fallen bald.
If you my opinion hear want, the stocks go-down soon.

b. Wenn du meine Meinung hören willst, fallen die Aktien bald.
If you my opinion hear want, go-down the stocks soon.

c. Wenn du meine Meinung hören willst, dann fallen die Aktien
If you my opinion hear want, then go-down the stocks
bald.
soon.

'If you want to hear my point of view, the stocks will go down soon.'
(Köpcke and Panther 1989, (49))

Again, what seems crucial for the unconditional readings of examples (107) and (108) is not the word order in the main clause, but rather, I argue, the extent to which common sense supports the notion that the propositions or events in antecedent and conditional are conditionally independent.[23]

Summary. Summing up, I argue that a purely pragmatic approach is reasonable and necessary. I concede that there are ambiguous conditionals $A > C$ such as (102) in which A and C could either be conditionally related or unrelated, and that in those undecided cases integrative or non-integrative word order will help decide on the reading of the sentence. But it is not so that in all cases the syntax or prosody of a sentence uniquely forces the pragmatic interpretation.[24] It then remains to be demonstrated how a suitable notion of conditional independence can do the pragmatic work that I claim it does.

23. That sentences like (107c), (108c) get unconditional readings may be critical to the account of BCs advanced by Ebert et al. (2008), according to which the (English) proform *then* forces a conditional reading (see the paper for details).

24. This suggests that there is something like a lexicographic order of strength, so to speak, according to which evidence for or against an unconditional reading is featured in interpretation: first and foremost the pragmatic question whether propositions A and C are plausibly conditionally (in)dependent is assessed; where this is (relatively) undecided syntactic information disambiguates readings.

5.3.3 Conditional Independence

The idea to explain the non-conditional readings of BCs pragmatically is very simple. Take again Austin's example (65a):

(65a) There are biscuits on the sideboard if you want them.

My explanation in a nutshell is this: since normally we would not expect the truth or falsity of propositions

you want some (A) & there are biscuits on the sideboard (C)

to depend on one another, a speaker who felicitously asserts (65a) *must* believe in —or be willing to defend, or purport to believe in, or purport to be willing to defend, or ...— the unconditional truth of C.[25] To spell out this idea we have to make precise what it means for two propositions to be independent in some appropriate sense.

Conditional Independence. I suggest that the right kind of independence of propositions is epistemic — epistemic in the sense that it governs how agents change their beliefs about one proposition when they change their beliefs —if only hypothetically— in the other. In other words, although the truth values of A and C might be fixed, what matters for our concern is whether propositions are normally believed to depend on one another. From this point of view we can say that A and C are *conditionally independent* for an agent (in a given epistemic state) if a minimal change in the belief about A will *not* result in a change in the belief about C, and vice versa.[26]

25. I know of two brief occurrences of this idea in the literature on conditionals. When discussing the pragmatics of certain 'odd conditionals', as he calls them, Frank Veltman reasons that any (data-semantic) information state which (i) supports a BC $A > C$, (ii) and supports $\Diamond A \wedge \Diamond \overline{A}$, *must* also support $\Box C$, as long as, Veltman reasons in a short bracketed remark, we don't expect the speaker to be able to merely *make* C true *at will* (Veltman 1986, p. 163).

A similar idea is also reported on in a footnote of a paper by Geis and Lycan (1993) where it says: "[Robert Stalnaker] does not buy our distinction of kind between 'NCCs' [read: non-conditional conditionals] and 'genuine' conditionals, but maintains that our alleged NCCs are genuine conditionals which only implicate their consequents; in context, Ad [the addressee] knows that Sp [the speaker] would not be asserting the conditional in question unless Sp had the truth of its consequent as a ground. We are unsure how the relevant Gricean reasoning would go, and/but we shall not try to criticize Stalnaker's view until he has spelled it out in writing." (Geis and Lycan 1993, p. 55, footnote 17) As far as I can tell, the work that comes closest to implementing Stalnaker's proposal in writing is a paper by Swanson (2003).

26. The nature of epistemic uncertainty does not play a role here. The account applies to epistemic, predictive and even counterfactual conditionals, as we will see later.

Here is a first formal take on the notion of conditional independence, geared towards a strict implication analysis. Take a set W of possible worlds, propositions $A, C \subseteq W$ and an agent's epistemic state $\sigma \subseteq W$ of worlds held possible. We say that $\Diamond A$ is true iff $\sigma \cap A \neq \emptyset$. With this define that A and C are CONDITIONALLY INDEPENDENT (on σ) iff

$$\forall X \in \left\{A, \overline{A}\right\}, \forall Y \in \left\{C, \overline{C}\right\} \; : \text{ if } \Diamond X \text{ and } \Diamond Y \text{ then } \Diamond(X \cap Y). \tag{5.12}$$

This notion captures the idea that two propositions are conditionally independent for an agent just in case learning one proposition to be true or false (where this was not decided before) is not enough evidence to decide whether the other proposition is true of false (where this was not decided before).

DERIVING UNCONDITIONAL READINGS. We can now make our initial idea more precise and derive unconditional readings under a strict implication analysis. If the speaker utters $A > C$, we may infer that, if she spoke truthfully, her epistemic state σ is such that $\sigma \cap A \subseteq C$. But if we also assume that the speaker does not believe in a conditional relationship between A and C in the sense of (5.12), we derive that the speaker either believes in the falsity of A or the truth of C. This is so, because if $\Diamond A$ and $\Diamond\overline{C}$, then by conditional independence we have $\Diamond(A \cap \overline{C})$ which contradicts $\sigma \cap A \subseteq C$. Consequently, if we furthermore assume $\Diamond A$ —by non-triviality presupposition— we may conclude that the speaker actually believes C.

UNCONDITIONAL VARIETY. Before justifying this account in more detail, let us first settle the issue to which cases it should or should not apply. Obviously, not all conditionals with conditionally independent propositions have the unconditional reading that C is true. The conditionals in (109), which we could call 'monkey's uncle'-conditionals, all convey that the speaker disbelieves the antecedent, and they supposedly do so in virtue of *modus tollens* and the commonsense assumption that the speaker believes in the falsity of the consequent.

(109) a. If that's true, I'm a monkey's uncle.
b. I'll be hanged, if my abstract got accepted.
c. If you are an astronaut, then I am the Emperor of China.

The derivation of unconditional readings by conditional independence does not apply to these conditionals, but we also don't need conditional independence to account for the intuitively attested readings of these sentences either.

Conversely, not all conditionals with an unconditional reading need to derive this reading by appeal to conditional independence. In other words, there are conditionals with unconditional readings whose propositions are *not* conditionally independent. Examples of these are given in (110).[27]

(110) a. This match is wet. If you strike it, it won't light.

b. Bij gladheid wordt niet gestrooid.
In case of slipperiness be-PASSIVE not spread.
'When icy, this road will not be salted.'

In general propositions like

you strike this match (A) & it will not light ($\overline{C}$)

or

the ground is frozen (A) & this road will not be salted ($\overline{C}$)

are conditionally dependent in the sense that we do think that normally A is (something like) a necessary condition for C. The conditionals in (110) are of the form $A > \overline{C}$, so that by mere reasoning about normal courses of events we should conclude that $\overline{C}$ is true as such, unconditional on A. This case is then similar to CP-readings that arise from (reasoning about) world knowledge in the form of commonsense normality assumptions (see section 5.2.2).

Another nice class of examples of conditionals with unconditional readings but conditionally related propositions are the sentences in (111).[28]

(111) a. This is the best book of the month, if not the year.

b. Some if not all of my friends are metalheads.

These conditionals are of the form $\overline{A} > C$, but it is safe to additionally assume that it is commonly understood that $A \subseteq C$. Whence that these conditionals also convey the unconditional truth of their consequents despite conditionally related propositions.

A final class of conditionals that may also get unconditional readings in a rather trivial manner are *echoic conditionals* like in (112), at least if we assume that these really presuppose the truth of their antecedents (cf. Iatridou 1991; Haegeman 2003).[29]

27. Example (110b) is from a road sign in Amsterdam's Westerpark.

28. Examples of this kind were brought to my attention by Frank Veltman.

29. I am not convinced that echoic conditionals really presuppose the truth of their antecedents. To me it sometimes rather seems that the antecedent is pending in the 'negotiation zone' between conversationalists. But this is not overly important here.

(112) a. If she is so pretty, you should have *her* do your laundry, not me.

b. If you are so smart, it is curious why you are unable to get a job.

c. If the wine bottle is half-empty, you are a pessimist. (Noh 1998)

If the truth of A is presupposed, the utterance of the conditional $A > C$ trivially derives the truth of C by *modus ponens*.

VARIATIONS ON INDEPENDENCE. Where does the notion of conditional independence come from? Why is it justified to use it in the way we do? And how does it relate to other comparable notions of independence? First of all, conditional independence is provably equivalent to Lewis (1988)'s notion of *orthogonality of questions*. This was observed and spelled out by van Rooij (2007) who took the notion I am suggesting here to account for the strengthening of conditional presuppositions.

Moreover, it is easy to verify that conditional independence is the purely qualitative counterpart to standard *probabilistic independence*. Propositions A and C are PROBABILISTICALLY INDEPENDENT given a probability distribution $\Pr(\cdot)$ iff $\Pr(A \cap C) = \Pr(A) \times \Pr(C)$. If we equate the epistemic state σ of an agent with the support of the probability distribution $\Pr(\cdot)$ as usual so that $\sigma = \{w \in W \mid \Pr(w) \neq 0\}$, we can show that probabilistic independence entails conditional independence. First, we establish that if $\Pr(A \cap C) = \Pr(A) \times \Pr(C)$, then for arbitrary $X \in \{A, \overline{A}\}$ and $Y \in \{C, \overline{C}\}$ it holds that $\Pr(X \cap Y) = \Pr(X) \times \Pr(Y)$. From the three arguments needed, it suffices to give just one, as the others are similar. So assume that $\Pr(A \cap C) = \Pr(A) \times \Pr(C)$ and derive that $\Pr(A \cap \overline{C}) = \Pr(A) \times \Pr(\overline{C})$: $\Pr(A \cap \overline{C}) = \Pr(A) - \Pr(A \cap C) = \Pr(A) - (\Pr(A) \times \Pr(C)) = \Pr(A) \times (1 - \Pr(C)) = \Pr(A) \times \Pr(\overline{C})$. Next, assume that $\Pr(X \cap Y) = \Pr(X) \times \Pr(Y)$ and that $\Diamond X$ and $\Diamond Y$. That means that $\Pr(X), \Pr(Y) > 0$. Hence, $\Pr(X \cap Y) > 0$, which is just to say that $\Diamond(X \cap Y)$.

The converse, however, is not the case. Conditional independence does not entail probabilistic independence. It may be the case that proposition A is not enough (evidence, support, information) to decide whether C is true or false, but still learning that A is true, for instance, makes C more or less likely.

Conditional independence is, however, strictly weaker than the more standard notion of *logical independence* if this latter notion is relativized to an epistemic state. The normal definition renders A and C LOGICALLY INDEPENDENT iff

$$\forall X \in \{A, \overline{A}\}, \ \forall Y \in \{C, \overline{C}\} \ : \ X \cap Y \neq \emptyset.$$

This is equivalent to the following formulation: take a maximally ignorant information state $\sigma = W$ that does not rule out any possible world whatsoever; then say that A and C are logically independent (on σ) iff for all $X \in \{A, \overline{A}\}$, $Y \in \{C, \overline{C}\}$: $\Diamond(X \cap Y)$. But although logical independence can thus be related to an all-inclusive information state, for our purposes this notion is inadequate to deal with restrictions of σ: generalizing logical independence of A and C to arbitrary information states σ in the most straightforward way excludes that either A or C is believed true or false in σ. In other words, logical independence does not have a flawless 'positive fit' if applied straightforwardly to epistemic states: there are instances of intuitively independent propositions which are not logically independent on some epistemic states. Conditional independence is weak enough to circumvent this problem.

Still, it might be objected that conditional independence as defined above is actually too weak to capture our intuitions about independence properly, for it shares with probabilistic independence the counterintuitive trait that if a proposition A is believed true, then any proposition C is independent of A, even A itself. In other words, conditional independence does not have a flawless 'negative fit': there are intuitively dependent propositions which are conditionally independent on some states. Yet so far this was not a problem for the above derivation of unconditional readings because (i) we have reasoned only *from* independence, and not *towards* it, so to speak, and, more importantly even, (ii) we have only looked at conditionals so far for which it was feasible to assume that the speaker was uncertain about the antecedent A. For these cases, the given notion of conditional independence applies non-vacuously and does the desired work for us.

Counterfactual Biscuits. A problem surfaces, however, when we turn to another interesting class of conditionals with unconditional readings, namely subjunctive or at least partially subjunctive examples as in (113).[30]

(113) a. If you had needed some money, there was some in the bank.
(Johnson-Laird 1986, (51))

b. If you would have wanted a beer, there were some in the fridge.

30. Example (113b) was brought up by Nathan Klinedinst as a problem case for an early version of the present account that I presented at PALMYR-V in Paris June 2nd 2007. I'm very grateful for this critical observation and the discussion that ensued. Also, van Rooij (2007) acknowledged this problem with the notion in (5.12).

Both of these sentences also convey the unrestricted truth of their consequents, but interestingly the antecedents have subjunctive mood marking and express counterfactuality. I suggest to speak of COUNTERFACTUAL BISCUIT CONDITIONALS or CBCs for short.

As far as I can tell, CBCs have not received much attention in the literature so far.[31] This is remarkable, since CBCs are interesting and relevant to the linguist's concerns in a number of ways. Firstly, with a subjunctive antecedent and a standard indicative consequent, the examples in (113) are hybrids between subjunctive and indicative and as such suggest themselves as an interesting test case for a compositional theory of tense and mood marking in conditionals. Unfortunately, this issue is way beyond the scope of this thesis.

Secondly, it is apparent that CBCs are problematic for naïve conditional speech-act accounts. For cases like (113) the analogue to a conditional-assertion analysis would have to be a *counterfactual-assertion analysis* which is curiously implausible: whereas in case of a conditional assertion a reasonable speech act is performed at least when the antecedent is true, a counterfactual assertion would never make it to assertion status, when the antecedent is presupposed false. It is then entirely unclear how conditional assertion approaches, if naïvely construed, could reasonably extend to CBCs.

UNCONDITIONAL COUNTERFACTUALS. Yet again, I do not think that unconditional readings arise only for special counterfactuals that we could address as CBCs. There are plain counterfactuals —i.e. with subjunctive mood marking both in antecedent and consequent— that function exactly like a standard BC would and convey the unconditional and *actual* truth of the consequent. Take the following small example dialogue:

(114) a. Bonnie: Are you hungry?
b. Clyde: No, I'm not.
c. Bonnie: Ah, that's a shame.
d. Clyde: Why is that?
e. Bonnie: If you had been hungry, there would have been pizza in the fridge.

To my mind, the counterfactual in (114e) is certainly felicitous in this context and it clearly conveys that there *is* pizza in the fridge, and not that pizza

31. Scheffler (2008*b*) deals with CBCs sentences briefly. Moreover, McCawley (1996) and von Fintel (1999) mention Johnson-Laird's example (113a) as curious but do not enlarge on it.

would have miraculously materialized there if Clyde had been hungry.[32]

As far as my intuition goes, we can even vary the German word order (and possibly even drop in the proform *dann*), and still, in a context like (114), all the variants in (115a)–(115d) are not only felicitous but do convey the unconditional reading that pizza is in fact in the fridge; only the variants (115e) and (115f) with indicative main clauses seem truly unacceptable.[33,34]

(115) Wenn du Hunger gehabt hättest, ...
If you hunger have-PART-PERF have-KONJ-2 ...

a. ...es wäre noch Pizza im Kühlschrank gewesen.
...it be-KONJ-2 still pizza in the fridge be-PART-PERF.

b. ...wäre noch Pizza im Kühlschrank gewesen.
...be-KONJ-2 still pizza in the fridge be-PART-PERF.

c. ...dann wäre noch Pizza im Kühlschrank gewesen.
...then be-KONJ-2 still pizza in the fridge be-PART-PERF.

d. ...es ist noch Pizza im Kühlschrank.
...it be-IND still pizza in the fridge.

e. *...ist noch Pizza im Kühlschrank.
...be-IND still pizza in the fridge.

f. *...dann ist noch Pizza im Kühlschrank.
...then be-IND still pizza in the fridge.

CHALLENGES OF UNCONDITIONAL COUNTERFACTUALS. Counterfactuals with unconditional readings pose a challenge to accounts of unconditional readings based on properties of BCs. It is not entirely obvious how the accounts of Ebert et al. (2008) and Scheffler (2008*a*,*b*) could derive these unconditional readings without the additional help of a theory of the kind that I am defending here. For these accounts, an assertion of (114e) comes down to an

32. I believe that (114e) is felicitous and conveys that there is pizza in the fridge, but ultimately my argument does not depend on the perfect felicity of (114e) as long as we acknowledge that even if (114e) is slightly (or not so slightly) odd, it *is* understood to convey that there is pizza in the fridge in a charitable conversation.

33. It is maybe advisable to compare the intuitive acceptability of the sentences in (115): even if the reader doubts the judgement that (115a)–(115d) are felicitous in a context like (114), the contrast remains between (115a)–(115d) on the one hand, and (115e) and (115f) on the other hand, the latter of which are clearly more marked.

34. The abbreviations PART-PERF, KONJ-2 and IND in the glosses for example (115) stand for "Partizip Perfekt" (past participle), "Konjunktiv 2" (roughly: subjunctive mood marker) and "Indikativ" (indicative).

assertion of the consequent (116) — supposing that this is the speech-act associated with the consequent in this context.

(116) There would have been pizza in the fridge.

An assertion of (116), however, does not directly establish that there *is*, but only that there *would be,* pizza in the fridge. The problem is that normally an assertion of a modalized expression "*would C*" does not flatly assert that *C* is the case. It can convey this meaning, of course. But the question is when exactly it does so and when exactly it does not. So, an explanation of unconditional readings as unconditional assertions of their consequents, as offered by Ebert et al. (2008) and Scheffler (2008*a*,*b*), though not falsified by this data, does not as such yet fully account for the unconditional readings of examples like (114e).

Intuitively, the idea of conditional independence does explain these cases just as well as the indicative cases we looked at before. There is no reasonable conditional relationship between the propositions

$$\text{you are hungry } (A) \qquad \& \qquad \text{there is pizza in the fridge } (C)$$

even when it is common ground that A is false: if we adopted the most conservative *counterfactual belief* in A we would not change our mind with respect to C. A similar reasoning as before should then yield that the only way of linking conditionally independent propositions in a counterfactual conditional is that the consequent must actually be true.[35]

This also explains why in some contexts a statement "*would C*" such as (116) can convey the actual truth of *C*: we may assume that the modal "*would*" is restricted by modal subordination to certain counterfactual worlds, say the most natural worlds where A is true, so that the proposition expressed is ultimately the same as that expressed by a counterfactual $A > C$ as in (114e); but then the same account for the derivation of unconditional readings can apply to the explicit counterfactual in (114e), as well as to the contextually restricted (116).

35. This can also be implemented in a semantic theory of counterfactuals that spells out laws and law-like connections explicitly and derives from this an ordering $\preceq_w$ on accessible worlds (see Veltman 2005; Schulz 2007). That is to say that the notion of conditional independence I suggest here should be regarded as a general interpretation constraint that some theories find easier to accommodate than others. The question remains whether a notion of independence cannot in some sense be *reduced* to properties of laws and facts alone. Such a reduction depends on the representation of laws and facts, of course, and is, as far as I can see, not trivial.

Generalizing Independence. Though, perhaps, intuitively appealing, my argument from conditional independence does not yet formally derive unconditional readings for counterfactuals. The formulation of conditional independence in (5.12) was matched to strict implication. For counterfactuals (and other kinds of conditionals) we would like to generalize the notion of independence to be compatible with order-sensitive implication.

Towards this end, we need to make the notion of independence sensitive to the ordering information represented in the modal structure $\langle R_w, \preceq_w \rangle$. The intuition behind the notion of conditional independence remains unchanged. We still say that A and C are *conditionally independent* for an agent (in a given epistemic state as represented by $\langle R_w, \preceq_w \rangle$) if a minimal change in the belief about A will *not* result in a change in the belief about C, and vice versa. In this spirit, say that C is conditionally independent of A (on a modal structure $\langle R_w, \preceq_w \rangle$) iff

$$\forall X \in \{A, \overline{A}\}, \ \forall Y \in \{C, \overline{C}\} \ : \ \Diamond Y \text{ iff } X \Diamond\!\!\Rightarrow Y.^{36} \tag{5.13}$$

This notion straightforwardly captures the intuition that C is independent of A if learning A does not change an agent's initial opinion as to whether C. Obviously, A and C are conditionally independent iff A is conditionally independent of C and C is conditionally independent of A.

This notion is in part a conservative extension of and in part an improvement of the previous formulation in (5.12). If we set $\sigma = \text{Min}_w$, then independence in the sense of (5.13) entails independence in the sense of (5.12). The reverse is not generally true. This is where the notion in (5.13) improves on the previous one in (5.12). Remember that according to (5.12), if an agent has a fixed belief in a proposition A, i.e., if $\Box A$ or $\Box\overline{A}$ is true on information state σ, then any proposition C is conditionally independent of A on σ in the sense of (5.12). The notion in (5.13), on the other hand, allows such 'circumstantial beliefs' not to interfere with the definition of independence, because it extends, so to speak, beyond Min_w in comparing beliefs in A and C.

It is still straightforward to show that independence as defined in (5.13) also successfully derives unconditional readings for indicatives if we apply an order-sensitive analysis to these. We would like to show that $A \Box\!\!\Rightarrow C$ implies

36. Recall that for all $X, Y \subseteq W$ we have:

$$\begin{aligned} \Diamond Y \quad &\text{iff} \quad \text{Min}_w \cap Y \neq \emptyset \\ X \Diamond\!\!\Rightarrow Y \quad &\text{iff} \quad \text{Min}_w(X) \cap Y \neq \emptyset. \end{aligned}$$

$\Box C$ if A and C are conditionally independent no matter what properties R_w and $\preceq_w$ have. This is indeed so, because if $\Diamond\overline{C}$ was true, we could derive that $A \mathrel{\Diamond\!\!\Rightarrow} \overline{C}$ was true from (5.13) which rules out that $A \mathrel{\Box\!\!\Rightarrow} C$ could be true.

Independent Counterfactuals. Moreover, the revised formulation of conditional independence also does some new work for us and helps account for the unconditional readings of counterfactuals. To deal with counterfactuals we would like a modal structure to represent information about an agent's disposition to revise her beliefs. Towards this end, let us assume that the ordering $\preceq_w$ represents similarity in the sense of Lewis (1973). More concretely, let $R_w = W$ contain *all* the possible worlds and let the relation $\preceq_w$ satisfy weak centering. It is then a simple argument that shows that if A and C are conditionally independent on $\langle R_w, \preceq_w \rangle$ in the sense of (5.13), and if the conditional $A > C$ is true in w, i.e., if $\mathrm{Min}_w(A) \subseteq [\![C]\!]$, then C is true in the actual world w. Above, we have already derived $\Box C$ from these conditions, which means that $\mathrm{Min}_w \subseteq C$. But then it suffices to note that weak centering guarantees that the actual world w is in Min_w and hence must make C true. This then derives the unconditional meaning of a conditional with counterfactual antecedent, no matter whether the main clause is in the subjunctive or the indicative.

5.3.4 Biscuits in Discourse

What is left to be explained is why a conditional with an unconditional reading should be used at all in conversation, given that its discourse effect, as far as information is concerned, is that of a simple assertion of the consequent. What purpose does the antecedent serve in an 'unconditional conditional'?

The Received View. The received view on the matter, found implicitly or explicitly in a lot of work on BCs in one form or another, appears to be something like this: the antecedent of a BC gives the conditions under which the speech act associated with C is felicitous (according to the speaker). According to the received view, the speaker is unsure whether a straightforward utterance of C would be felicitous, but believes that A is (likely enough) a sufficient condition for a felicitous utterance of C. Hedging the statement by uttering a conditional $A > C$ instead of a plain use of C then makes (sufficiently) sure that, as far as the speaker is concerned, the whole utterance is felicitous.

Biscuits as Interpretation Cues. I would like to argue that the received view, as I have spelled it out here, is mistaken. It is not the *truth* of the antecedent that serves to establish felicity, but rather it is the mere *use* of the antecedent, *the fact that it was produced* that helps assure felicity. I will argue towards this conclusion based on our intuitions about two situated examples.

Here is my first example. Imagine that we want to go swimming and you are waiting for me while I am packing my bag. If I now say to you —out of the blue— that

(117) There are biscuits on the sideboard (*C*).

it is conceivable, if not likely that you may not know what exactly I meant to tell you (cf. Cappelen and Lepore 2005, on speech-act pluralism): May you eat the biscuits? Do I want you to stay away from them? Must you hand them to me? Throw them into my bag? You may be unsure, even though you are *in fact* hungry and lust for sweets and I know it. The critical point is that it may not be intelligible *in which way* the utterance of *C* has to be understood, maybe because it is not *common ground* that you would like to eat biscuits, although this is true and known by both speaker and hearer. In contrast, the Austinean bc in (65a) makes entirely clear for what reason the information *C* is given. This example suggests that the function of the antecedent is to make an utterance of *C* intelligible, to help understand *how* the information *C* has to be treated and processed.

Here is another similar example that makes a related but slightly different point. In certain contexts, different antecedents may change the interpretation of the consequent dramatically. Just compare the sentence (118a) from the quote that opened this chapter, with the sentence in (118b) that notably has the exact same consequent.

(118) a. If you need anything, I'm Jill.
b. If you want to go out tonight, I'm Jill.

Though used in the same context of utterance, the interpretation of the consequent *C* differs substantially: sentence (118a) might encourage the hearer to ask for help (as a customer), while the sentence (118b) might encourage him to ask for the speaker's phone number (or some such). Again, the example shows how the antecedent may specify or disambiguate the interpretation of the consequent, i.e., how it affects its broader integration, reception and processing in discourse.

Taken together, these two examples support the idea that it is not necessarily the case that the *truth* of the antecedent guarantees felicity and relevance

of the speech act associated with the consequent, but rather that *giving* the antecedent contributes to relevance or felicity.

This brings up a discourse function of conditionals that has so far not been explicitly discussed in the literature, as far as I can tell. I propose to think of *some* conditionals as '*intelligibility conditionals*': the antecedent is given to cue the proper reception and interpretation of the consequent. This is certainly what is going on in (119a) and plausibly also in (119b).[37]

(119) a. He's a buhubahuba, if you know what I mean.

b. He trapped two mongeese, if that's how you make the plural of "mongoose." (Noh 1998)

Not all BCs are intelligibility conditionals in this sense: witness, for instance, politeness-hedging and speaker-attitude commenting with "if I may say so," "if you ask me," "if I'm honest," "if I may interrupt" etc.

Context-Shifts for Optimality. The game theoretic model that we used to explain conditional perfection readings in section 5.2.4 neatly captures such discourse functions of conditionals. Section 5.2.4 introduced the idea of an extensive game with dynamic unawareness in order to explain how a conditional $A > C$ is to be interpreted as an answer to the contextual question whether C is true. I argued that the conditional shifts the interpreter's conceptualization of the context of utterance: while a simple assertion of C would have been interpreted in the light of a simple game $G_{?C}$, the conditional forces the hearer to revise his conception of the signaling game to a more complex game $G^{+}_{?C}$.

A similar process of comparison between a simple signaling game G and a revised game G^{+} also explains the discourse function of BCs and related constructions. If the interpreter derives an unconditional reading from independence, forming the belief that C is true, it is natural to compare the utterance of $A > C$ with a simple utterance of C. Thus conceived it is the presence of the antecedent, not its truth, which helps establish felicity, be that in the form of relevance or intelligibility.

Moreover, just as in the case of conditional perfection under a contextual question $?C$, the use of the conditional is to be rationalized *ex post*. We predict some sort of conditional perfection, in a loose manner of speaking perhaps,

37. In example, (119b) the speaker might either worry about not being understood, about saying something ungrammatical (while still being understood), or both.

also for conditionals with unconditional readings. For intelligibility conditionals the hearer could conclude that the speaker was not sufficiently sure that the simple utterance *C* would have been interpreted appropriately. For other kinds of BCs the perfection inference here would be different: indeed it may be that the hearer comes to believe that the speaker thought that an utterance of *C* would have been impolite, ill-formed or otherwise infelicitous. At the heart of this explanation is the idea that forward induction reasoning naturally models language interpretation as rationalization in an *ex post* constructed context.

5.3.5 Projection and a Big Fat Lie

I would like to conclude the discussion of unconditional readings by a defense of my account against possible criticism based on certain aberrant examples that were featured prominently in the recent discussion of BCs. To begin with, consider the following examples of BCs:

(120) a. (The door bell is ringing.)
Mary to Jane: If that's John, I'm not here. (Noh 1998, (65))

b. If anyone talks to you about the treasure map, you don't know anything about it, you have never heard of it. (Noh 1998, (66))

c. If they ask you how old you are, you're four. (Siegel 2006, (8))

The examples in (120) are special in that their antecedents should not be taken as flat, honest and credible assertions, but rather as *directives*: intuitively, the speaker urges the hearer into performing a certain action, in particular, into behaving as if *C* was true in (at least) those situations in which *A* is true. We could speak of these as *projections* in the sense that the speaker projects onto the hearer commitment to the truth of the consequent (in a certain sense), rather than to believe it, or be willing to defend it herself. But then, shouldn't these projection examples be problematic for the account that I have given here? After all, the account given here derives that the speaker believes that the consequent is true.

The same worry arises in connection with the following example:[38]

(121) If you want to hear a big fat lie, George W. and Condi Rice are secretly married. (Siegel 2006, (22))

38. I am grateful to Cornelia Ebert for raising this issue.

Indeed, as Ebert et al. (2008) rightly point out, the speech act associated with the consequent in (121) "cannot be a run-of-the-mill assertion since it has been explicitly classified as a lie beforehand." Certainly, it also seems dubious to claim that the speaker believes the consequent of (121).

Still, I do not think that either the projection cases (120), nor the 'big fat lie' in (121) prove my account of unconditional readings wrong. The point is simply that I am not committed to the assumption that pragmatic reasoning stops once it has established that an utterance normally conveys that the speaker believes such and such. Irony and sarcasm most likely also start with a literal interpretation: indeed, one of the main ideas of the IBR model is that literal and credulous interpretation is a natural starting point that can be overthrown by further pragmatic consideration. So, I don't think it is implausible at all to maintain that the derivation of an unconditional reading could proceed as sketched above, but that the hearer continues to interpret, roughly, as follows: so I should conclude that the speaker believes that C, but that is not plausible (because she certainly knows that not C) and she probably rather *does as if* she believes C in order for me to realize that (i) she wants me to behave as if C was true (in certain confined circumstances; for *her* benefit, etc.), or (ii) she wants me to entertain the untrue thought that C is true (and that she thinks that this is hilariously funny).

SUMMARY. Let me then briefly sum up this chapter. I have argued for a contextualist treatment of conditional perfection and unconditional readings of conditionals: in many cases commonsensical assumptions about the context of utterance derive these readings. This is not always the case. Some conditional perfection readings require a more genuinely pragmatic explanation in terms of reasoning about the available alternative answers to a topical question under discussion. Similarly, not all unconditional readings need to be derived by appeal to conditional independence. Still, intuitions about conditional relatedness are strong enough to even overrule cues from word order that have been taken as constitutive of the class of biscuit conditionals.

Chapter 6

Conclusions & Outlook

> What we call the beginning is often the end
> And to make an end is to make a beginning.
> The end is where we start from. [...]
> We shall not cease from exploration
> And the end of all our exploring
> Will be to arrive where we started
> And know the place for the first time. (T.S. Eliot, *Little Gidding*)

In the preceding five chapters I have spelled out a model of step-by-step pragmatic reasoning that incorporates introspectively plausible and empirically vindicated assumptions about the psychology of reasoners. The resulting IBR model is a refinement of rationalizability in which additional assumptions about the cognitive architecture of language users are implemented explicitly in the belief formation process of agents. By additionally giving general principles for the construction and interpretation of signaling games as models of the context of utterance, I have shown how the IBR model accounts for a variety of data such as embedded scalar implicatures, free choice readings and the like. The model's explicit epistemic approach offered a novel perspective on the interpretation of bidirectional optimality theory and proved helpful in characterizing the development of pragmatic competence surrounding scalar implicatures in early acquisition.

Unsurprisingly, not all questions have been answered; hopefully, some have; probably, some old questions appear still unanswered in a new clearer light; and, certainly, some new questions surfaced for future consideration. Let me just point out some of the most pressing issues here, some of which have and some of which have not been addressed in the text so far.

For one, although chapter 2 provided some crucial insight into the formal characteristics of the IBR model, more results of the same sort would be

welcome. Chapter 2 ended with the conjecture about a proper epistemic characterization result of IBR as a solution concept. In future work, I would be curious to test this conjecture by giving a full epistemic characterization. Similarly, I would appreciate an answer to the question whether there is a natural class of signaling games for which the IBR model always reaches a fixed point, and perhaps even the same fixed point for both sequences.

Another open issue is an interpretation of the IBR model as a model of language change. As mentioned in section 2.4.2, formally speaking the IBR model *as is* could be taken as a diachronic model implementing a special form of *best-response dynamics*. This would allow many further applications and would also allow further comparison to bidirectional optimality theory where the latter is considered a diachronic model too. To justify the use of the IBR model as a diachronic model, however, it would be necessary to go through the set of assumptions that informed its present formulation, all of which were motivated by appeal to empirically or intuitively reasonable assumptions about human reasoning. It would then be essential to see whether and how these assumptions can be brought to bear on a model of language evolution. Related but in a sense orthogonal to this project is to check whether the IBR model as a model of individual reasoning could not be combined with existing models of learning and diachronic adaptation. Both of these issues seem very promising and interesting topics for future research.

Finally, I believe that an extension of the IBR model that incorporates reasoning about unawareness, as briefly introduced in section 5.2.4, could be very fruitfully applied to matters of linguistic and philosophical interest. For instance, Sperber and Wilson (1995) argue that certain features of a conversational context should not be considered beliefs of an agent, but be subjected to a different, weaker epistemic relation which they call MUTUAL MANIFESTNESS (Sperber and Wilson 1995, p. 38–46). Similarly, I have argued against a standard interpretation of prior probabilities in game models in section 3.1 as specifications of hearer beliefs. I suggested that prior probabilities in context models are best conceived of as a condensed representation of the associative strength with which an interpretation comes to mind when hearing a given form. This still leaves many questions open, but I have the hunch that including the dynamics of awareness in relevant ways into our game model could solve many outstanding issues with classical beliefs and probabilities. This may eventually lead to interesting new insights concerning issues such as transparency and recognition of the speaker's communicative intention.

References

Allott, Nicholas (2006). "Game Theory and Communication". In: *Game Theory and Pragmatics*. Ed. by Anton Benz et al. Palgrave MacMillan. Pp. 123–151.

Aloni, Maria (2007). "Expressing Ignorance or Indifference. Modal Implicatures in Bi-Directional Optimality Theory". In: *Logic, Language and Computation: Papers from the 6th International Tbilisi Symposium*. Ed. by Balder ten Cate and Henk Zeevat. Vol. 4363. Berlin: Springer Verlag. Pp. 1–20.

Aloni, Maria and Robert van Rooij (2007). "Free Choice Items and Alternatives". In: *Cognitive Foundations of Interpretation*. Ed. by Gerlof Bouma et al. Amsterdam: KNAW. Pp. 5–25.

Alonso-Ovalle, Luis (2005). "Distributing the Disjuncts over the Modal Space". In: *Proceedings of the North East Linguistics Society*. Ed. by Leah Bateman and Cherlon Ussery. Vol. 35. GLSA. Amherst, MA.

Anscombre, Jean-Claude and Oswald Ducrot (1983). *L'argumentation dans la langue*. Brussels: Mardaga.

Asher, Nicholas and Daniel Bonevac (2005). "Free Choice Permission as Strong Permission". In: *Synthese* 145.3. Pp. 303–323.

Asher, Nicholas et al. (2001). "Game Theoretical Foundations for Gricean Constraints". In: *Proceedings of the 13th Amsterdam Colloquium*. Ed. by Robert van Rooy and Martin Stokhof. Pp. 31–37.

Atlas, Jay David and Stephen Levinson (1981). "It-clefts, Informativeness, and Logical Form". In: *Radical Pragmatics*. Ed. by Peter Cole. Academic Press. Pp. 1–61.

Aumann, Robert (1974). "Subjectivity and Correlation in Randomized Strategies". In: *Journal of Mathematical Economics* 1.1. Pp. 67–96.

Aumann, Robert (1995). "Backward Induction and Common Knowledge of Rationality". In: *Games and Economic Behavior* 8.1. Pp. 6–19.

Aumann, Robert and Adam Brandenburger (1995). "Epistemic Conditions for Nash Equilibrium". In: *Econometrica* 63.5. Pp. 1161–1180.

Austin, John L. (1956). "Ifs and Cans". In: *Proceedings of the British Academy* 42. Pp. 109–132.

van der Auwera, Johan (1986). "Conditionals and Speech Acts". In: *On Conditionals*. Ed. by Elizabeth Closs Traugott et al. Cambridge University Press. Pp. 197–214.

— (1997*a*). "Conditional Perfection". In: *On Conditionals Again*. Ed. by A. Athanasiadou and R. Dirven. John Benjamins. Pp. 169–190.

— (1997*b*). "Pragmatics in the Last Quarter Century: The Case of Conditional Perfection". In: *Journal of Pragmatics* 27.3. Pp. 261–274.

Bach, Kent (1994). "Semantic Slack: What is Said and More". In: *Foundations of Speech Act Theory. Philosophical and Linguistic Perspectives*. Ed. by Savas L. Tsohatzidis. London and New York: Routledge. Pp. 267–291.

— (1999). "The Myth of Conventional Implicature". In: *Linguistics and Philosophy* 22.4. Pp. 327–366.

Banks, Jeffrey S. and Joel Sobel (1987). "Equilibrium Selection in Signaling Games". In: *Econometrica* 55.3. Pp. 647–661.

Banks, Jeffrey S. et al. (1994). "An Experimental Analysis of Nash Refinements in Signaling Games". In: *Games and Economic Behavior* 6.1. Pp. 1–31.

Bar-Hillel, May and Ruma Falk (1982). "Some Teasers Concerning Conditional Probabilities". In: *Cognition* 11.2. Pp. 109–122.

Basu, Kaushik and Jörgen W. Weibull (1991). "Strategy Subsets Closed under Rational Behavior". In: *Economic Letters* 36.1991. Pp. 141–146.

Battigalli, Pierpaolo (1996). "Strategic Rationality Orderings and the Best Rationalization Principle". In: *Games and Economic Behavior* 13. Pp. 178–200.

— (2006). "Rationalization in Signaling Games: Theory and Applications". In: *International Game Theory Review* 8.1. Pp. 67–93.

Battigalli, Pierpaolo and Marciano Siniscalchi (2002). "Strong Belief and Forward Induction Reasoning". In: *Journal of Economic Theory* 106. Pp. 356–391.

Beaver, David and Hanjung Lee (2004). "Input-Output Mismatches in Optimality Theory". In: *Optimality Theory and Pragmatics*. Ed. by Reinhard Blutner and Henk Zeevat. Palgrave MacMillan. Chap. 6, pp. 112–153.

Ben-Porath, Elchanan and Eddie Dekel (1992). "Signaling Future Actions and the Potential for Sacrifice". In: *Journal of Economic Theory* 57. Pp. 36–51.

Bennett, Jonathan (2003). *A Philosophical Guide to Conditionals*. Oxford University Press.

Benz, Anton (2006). "Utility and Relevance of Answers". In: *Game Theory and Pragmatics*. Ed. by Anton Benz et al. Palgrave. Pp. 195–219.

— (2007). "On Relevance Scale Approaches". In: *Proceedings of Sinn und Bedeutung 11*. Ed. by Estela Puig-Waldmüller. Pp. 91–105.

— (2009). "How to Set Up Normal Optimal Answers Models". Unpublished manuscript, ZAS Berlin.

Benz, Anton and Robert van Rooij (2007). "Optimal Assertions and what they Implicate". In: *Topoi* 26. Pp. 63–78.

Benz, Anton et al., eds. (2006). *Game Theory and Pragmatics*. Hampshire: Palgrave MacMillan.

Bernardo, José M. (1979). "Expected Information as Expected Utility". In: *The Annals of Statistics* 7.3. Pp. 686–690.

Bernheim, B. Douglas (1984). "Rationalizable Strategic Behavior". In: *Econometrica* 52.4. Pp. 1007–1028.

Bhatt, Rajesh and Roumyana Pancheva (2006). "Conditionals". In: *The Blackwell Companion to Syntax*. Vol. 1. Blackwell. Chap. 16, pp. 638–687.

Block, Eliza (2008). "Is the Symmetry Problem Really a Problem?" Unpublished manuscript, NYU.

Blume, Lawrence et al. (1991*a*). "Lexicographic Probabilities and Choice Under Uncertainty". In: *Econometrica* 59.1. Pp. 61–79.

— (1991*b*). "Lexicographic Probabilities and Equilibrium Refinements". In: *Econometrica* 59.1. Pp. 91–98.

Blutner, Reinhard (1998). "Lexical Pragmatics". In: *Journal of Semantics* 15. Pp. 115–162.

— (2000). "Some Aspects of Optimality in Natural Language Interpretation". In: *Journal of Semantics* 17. Pp. 189–216.

Blutner, Reinhard and Henk Zeevat, eds. (2004). *Optimality Theory and Pragmatics*. Palgrave MacMillan.

— (2008). "Optimality-Theoretic Pragmatics". To appear in: Claudia Maienborn, Klaus von Heusinger and Paul Portner (eds.) *Semantics: An International Handbook of Natural Language Meaning*.

Boër, Steven E. and William G. Lycan (1973). "Invited Inferences and Other Unwelcome Guests". In: *Papers in Linguistics* 6. Pp. 453–506.

Borg, Emma (2004). *Minimal Semantics*. Oxford University Press.

Bott, Lewis and Ira A. Noveck (2004). "Some Utterances are Underinformative: The Onset and Time Course of Scalar Inferences". In: *Journal of Memory and Language* 51.3. Pp. 437–457.

Breheny, Richard and Napoleon Katsos (2008). *Experimental Investigations of the Semantics/Pragmatics Interface*. Course Material for ESSLLI.

Breheny, Richard et al. (2006). "Are Generalised Scalar Implicatures Generated by Default? An On-Line Investigation into the Role of Context in Generating Pragmatic Inferences". In: *Cognition* 100.3. Pp. 434–463.

Camerer, Colin F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Camerer, Colin F. et al. (2004). "A Cognitive Hierarchy Model of Games". In: *The Quarterly Journal of Economics* 119.3. Pp. 861–898.

van Canegem-Ardijns, Ingrid and William van Belle (2008). "Conditionals and Types of Conditional Perfection". In: *Journal of Pragmatics* 40. Pp. 349–376.

Cappelen, Herman and Ernie Lepore (2005). *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Oxford: Blackwell.

Carston, Robyn (1988). "Implicature, Explicature, and Truth-Theoretic Semantics". In: *Mental Representations: The Interface Between Language and Reality*. Ed. by Ruth M. Kempson. Cambridge: Cambridge University Press. Pp. 155–181.

— (1998). "Informativeness, Relevance and Scalar Implicature". In: *Relevance Theory: Applications and Implications*. Ed. by Robyn Carston and S. Uchida. Amsterdam: John Benjamins. Pp. 179–236.

Chapman, Siobhan (2005). *Paul Grice, Philosopher and Linguist*. Hampshire: Palgrave MacMillan.

Chierchia, Gennaro (2004). "Scalar Implicatures, Polarity Phenomena and the Syntax/Pragmatics Interface". In: *Structures and Beyond*. Ed. by Adriana Belletti. Oxford University Press. Pp. 39–103.

Chierchia, Gennaro et al. (2008). "The Grammatical View of Scalar Implicatures and the Relationship between Semantics and Pragmatics". Unpublished manuscript.

Cho, In-Koo and David M Kreps (1987). "Signaling Games and Stable Equilibria". In: *The Quarterly Journal of Economics* 102.2. Pp. 179–221.

Clark, Herbert H. and Catherine R. Marshall (1981). "Definite Reference and Mutual Knowledge". In: *Elements of Discourse Understanding*. Ed. by Aravind K. Joshi et al. Cambridge University Press. Pp. 10–63.

Colman, Andrew M. (2003). "Depth of Strategic Reasoning in Games". In: *Trends in Cognitive Sciences* 7.1. Pp. 2–4.

Comrie, Bernard (1986). "Conditionals: A Typology". In: *On Conditionals*. Ed. by Elizabeth Closs Traugott et al. Cambridge University Press. Pp. 77–99.

de Cornulier, Benoît (1983). "'If' and the Presumption of Exhaustivity". In: *Journal of Pragmatics* 7. Pp. 247–249.

Costa-Gomes, Miguel A. et al. (2009). "Comparing Models of Strategic Thinking in van Huyck, Battalio and Beil's Coordination Games". In: *Journal of the European Economic Association* 7.2–3. Pp. 365–376.

Crawford, Vincent P. (2003). "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions". In: *American Economic Review* 93.1. Pp. 133–149.

Crawford, Vincent P. and Nagore Iriberri (2007). "Fatal Attraction: Salience, Naïveté, and Sophistication in Experimental "Hide-and-Seek" Games". In: *The American Economic Review* 97.5. Pp. 1731–1750.

Culicover, Peter W. and Ray Jackendoff (1997). "Semantic subordination despite syntactic coordination". In: *Linguistic Inquiry* 28.2. Pp. 195–217.

van Damme, Eric (1989). "Stable Equilibria and Forward Induction". In: *Journal of Economic Theory* 48. Pp. 476–469.

Dancygier, Barbara (1998). *Conditionals and Predictions. Time, Knowledge and Causation in Conditional Constructions*. Cambridge University Press.

Davidson, Donald (1974). "Belief and the Basis of Meaning". In: *Synthese* 27.3–4. Pp. 309–323.

— (1986). "A Nice Derangement of Epitaphs". In: *Truth and Interpretation*. Ed. by Ernest LePore. Oxford: Blackwell. Pp. 433–446.

Davison, Alice (1983). "Linguistic or Pragmatic Description in the Context of the Performadox". In: *Linguistics and Philosophy* 6.4. Pp. 499–526.

Dekker, Paul and Robert van Rooij (2000). "Bi-Directional Optimality Theory: An Application of Game Theory". In: *Journal of Semantics* 17. Pp. 217–242.

DeRose, Keith and Richard E. Grandy (1999). "Conditional Assertions and 'Biscuit' Conditionals". In: *Noûs* 33.3. Pp. 405–420.

van Dijk, Teun A. (1979). "Pragmatic Connectives". In: *Journal of Pragmatics* 3. Pp. 447–456.

Ducrot, Oswald (1969). "Présupposés et Sous-Entendus". In: *Langue Français* 4. Pp. 30–43.

— (1973). *La preuve et le dire*. Maison Mame.

Ebert, Christian et al. (2008). "A Unified Analysis of Indicative and Biscuit Conditionals as Topics". To appear in Proceedings of SALT XVIII.

Edgington, Dorothy (1995). "On Conditionals". In: *Mind* 104.413. Pp. 235–329.

Falk, Ruma (1992). "A Closer Look at the Probabilities of the Notorious Three Prisoners". In: *Cognition* 43.3. Pp. 197–223.

Farrell, Joseph (1988). "Communication, Coordination and Nash Equilibrium". In: *Economic Letters* 27.3. Pp. 209–214.

— (1993). "Meaning and Credibility in Cheap-Talk Games". In: *Games and Economic Behavior* 5. Pp. 514–531.

Farrell, Joseph and Matthew Rabin (1996). "Cheap Talk". In: *The Journal of Economic Perspectives* 10.3. Pp. 103–118.

Feinberg, Yossi (2004). "Subjective Reasoning — Games with Unawareness". Research Paper No. 1875, Stanford University.

— (2005). "Games with Incomplete Awareness". Research Paper No. 1894, Stanford University.

Fine, Kit (1975). "Critical Notice: Counterfactuals." In: *Mind* 84.335. Pp. 451–458.

von Fintel, Kai (1999). "The Presupposition of Subjunctive Conditionals". In: *The Interpretive Tract*. Ed. by Uli Sauerland and Orin Percus. MIT Working Papers in Linguistics 25. MIT. Cambridge, Massachusetts. Pp. 29–44.

— (2001*a*). "Conditional Strengthening – A Case Study in Implicature". Unpublished manuscript, MIT.

— (2001*b*). "Counterfactuals in a Dynamic Context". In: *Ken Hale: A Life in Language*. Ed. by Michael Kenstowicz. MIT Press. Pp. 123–152.

Flobbe, Liesbeth et al. (2008). "Children's Application of Theory of Mind in Reasoning and Language". In: *Journal of Logic, Language and Information* 17. Pp. 417–442.

Fox, Craig R. and Jonathan Levav (2004). "Partition-Edit-Count: Naive Extensional Reasoning in Judgement of Conditional Probability". In: *Journal of Experimental Psychology: General* 133.4. Pp. 626–642.

Fox, Danny (2007). "Free Choice and the Theory of Scalar Implicatures". In: *Presupposition and Implicature in Compositional Semantics*. Ed. by Ulrich Sauerland and Penka Stateva. Hampshire: Palgrave MacMillan. Pp. 71–120.

van Fraassen, Bas (1973). "Values and the Heart's Command". In: *Journal of Philosophy* 70.1. Pp. 5–19.

Franke, Michael (2008*a*). "Interpretation of Optimal Signals". In: *New Perspectives on Games and Interaction*. Ed. by Krzysztof R. Apt and Robert van Rooij. Vol. 4. Texts in Logic and Games. Amsterdam University Press. Pp. 297–310.

— (2008*b*). "Pseudo-Imperatives and Other Cases of Conditional Conjunction and Conjunctive Disjunction". In: *'Subordination' versus 'Coordination'*

in Sentence and Text — From a Cross-Linguistic Perspective. Ed. by Cathrine Fabricius-Hansen and Wiebke Ramm. Studies in Language Companion Series (SLCS). Amsterdam, Philadelphia: John Benjamins. Pp. 255–279.

Franke, Michael et al. (to appear). "Relevance in Cooperation and Conflict". To appear in Journal of Logic and Computation.

Fudenberg, Drew and David K. Levine (1998). *The Theory of Learning in Games*. MIT Press.

Gazdar, Gerald (1979). *Pragmatics: Implicature, Presupposition, and Logical Form*. New York: Academic Press.

Geis, Michael L. and William G. Lycan (1993). "Nonconditional Conditionals". In: *Philosophical Issues* 21.2. Pp. 35–56.

Geis, Michael L. and Arnold M. Zwicky (1971). "On Invited Inferences". In: *Linguistic Inquiry* 2.4. Pp. 561–566.

Gessen, Keith (2008). *All the Sad Young Literary Men*. Penguin Books.

Geurts, Bart (2005). "Entertaining Alternatives: Disjunctions as Modals". In: *Natural Language Semantics* 13. Pp. 383–410.

— (2009). "Scalar Implicature and Local Pragmatics". In: *Mind and Language* 24.1. Pp. 51–79.

Gibbons, Robert (1992). *A Primer in Game Theory*. New York: Harvester Wheatsheaf.

Gilboa, Itzhak and Akihiko Matsui (1991). "Social Stability and Equilibrium". In: *Econometrica* 59.3. Pp. 859–867.

Gómez-Txurruka, Isabell (2002). *The Semantics of Natural Language Disjunction Or*. Unpublished manuscript. ILCLI Donostia-San Sebastián.

Goodman, Nelson (1947). "The Problem of Counterfactual Conditionals". In: *Journal of Philosophy* 44.5. Pp. 113–128.

Grafen, Alan (1990). "Biological Signals as Handicaps". In: *Journal of Theoretical Biology* 144. Pp. 517–546.

Green, Mitchel S. (1995). "Quantity, Volubility, and some Varieties of Discourse". In: *Linguistics and Philosophy* 18.1. Pp. 83–112.

Grice, Paul Herbert (1989). *Studies in the Ways of Words*. Harvard University Press.

Groenendijk, Jeroen and Martin Stokhof (1984). "Studies in the Semantics of Questions and the Pragmatics of Answers". PhD thesis. Universiteit van Amsterdam.

Grünwald, Peter and Joseph Y. Halpern (2003). "Updating Probabilities". In: *Journal of Artificial Intelligence Research* 19. Pp. 243–278.

Gualmini, Andrea (2007). "Scope Resolution and Overt Questions: A Test for the QAR". In: *Proceedings of the Eighth Tokyo Conference on Psycholinguistics*. Ed. by Y. Otsu. Pp. 121–135.

— (2008). "The Rise and Fall of Isomorphism". In: *Lingua* 118.8. Pp. 1158–1176.

Günthner, Susanne (1999). "Wenn-Sätze im Vor-Vorfeld: Ihre Formen und Funktionen in der gesprochenen Sprache". In: *Deutsche Sprache* 3. Pp. 209–235.

Haegeman, Liliane (2003). "Conditional Clauses: External and Internal Syntax". In: *Mind and Language* 18.4. Pp. 317–339.

Halpern, Joseph Y. (2009). "Lexicographic Probability, Conditional Probability and Nonstandard Probability". To appear in *Games and Economic Behavior*.

Hanna, Joy E. et al. (2003). "The Effects of Common Ground and Perspective on Domains of Referential Interpretation". In: *Journal of Memory and Language* 49.1. Pp. 43–61.

Harsanyi, John C. (1967). "Games with Incomplete Information Played by 'Bayesian' Players, I–III: Part I. The Basic Model". In: *Management Science* 14.3. Pp. 159–182.

— (1968*a*). "Games with Incomplete Information Played by 'Bayesian' Players, I–III: Part II, Bayesian Equilibrium Points". In: *Management Science* 14.5. Pp. 320–334.

— (1968*b*). "Games with Incomplete Information Played by 'Bayesian' Players, I–III: Part III. The Basic Probability Distribution of the Game". In: *Management Science* 14.7. Pp. 486–502.

Heap, Shaun P. Hargreaves and Yanis Varoufakis (2004). *Game Theory — A Critical Text (Second Edition)*. Routledge.

Hedden, Trey and Jun Zhang (2002). "What Do You Think I Think You Think?: Strategic Reasoning in Matrix Games". In: *Cognition* 85.1. Pp. 1–36.

Heifetz, Aviad et al. (2009). "Dynamic Unawareness and Rationalizable Behavior". Unpublished manuscript.

Heller, Daphna et al. (2008). "The Role of Perspective in Identifying Domains of Reference". In: *Cognition* 108.3. Pp. 831–836.

Hendriks, Petra (2008). "A Unified Explanation for Production/Comprehension Asymmetries". In: *Proceedings of GALA 2007*. Ed. by Anna Gavarró Algueró and M. Joao Freitas. Cambridge Scholars Publishing. Pp. 240–251.

Hendriks, Petra and Helen de Hoop (2001). "Optimality Theoretic Semantics". In: *Linguistics and Philosophy* 24. Pp. 1–32.

Hendriks, Petra and Jennifer Spenader (2005). "When Production Precedes Comprehension: An Optimization Approach to the Acquisition of Pronouns". In: *Language Acquisition* 13.4. Pp. 319–348.

Hendriks, Petra et al. (2007). "Conflicts in Interpretation". Unpublished book manuscript, Groningen, Nijmegen, Utrecht.

Hintikka, Jaakko (1986). "Logic of Conversation as a Logic of Dialogue". In: *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. Ed. by Richard Grandy and Richard Warner. Clarendon. Pp. 259–276.

Ho, Teck-Hua et al. (1998). "Iterated Dominance and Iterated Best Response in Experimental 'p-Beauty Contests' ". In: *The American Economic Review* 88.4. Pp. 947–969.

Holt, Debra (1999). "An Empirical Model of Strategic Choice with an Application to Coordination Games". In: *Games and Economic Behavior* 27.1. Pp. 86–105.

de Hoop, Helen and Irene Krämer (2005). "Children's Optimal Interpretations of Indefinite Subjects and Objects". In: *Language Acquisition* 13.2. Pp. 103–123.

Horn, Laurence R. (1972). "On the Semantic Properties of Logical Operators in English". PhD thesis. UCLA.

— (1984). "Towards a New Taxonomy for Pragmatic Inference: Q-based and I-based implicatures". In: *Meaning, Form, and Use in Context*. Ed. by Deborah Shiffrin. Washington: Georgetown University Press. Pp. 11–42.

— (1989). *A Natural History of Negation*. Chicago: Chicago University Press.

— (2000). "From *if* it *iff*: Conditional Perfection as Pragmatic Strengthening". In: *Journal of Pragmatics* 32. Pp. 289–326.

— (2004). "Implicature". In: *The Handbook of Pragmatics*. Ed. by Laurence R. Horn and Gregory Ward. Blackwell. Pp. 3–28.

Hulsey, Sarah et al. (2004). "The Question-Answer Requirement and Scope Assignment". In: *Plato's Problems: Papers on Language Acquisition*. Ed. by Aniko Csirmaz et al. Vol. 48. MIT Working Papers in Linguistics. Department of Linguistics and Philosophy, MIT. Pp. 71–90.

Iatridou, Sabine (1991). "Topics in Conditionals". PhD thesis. MIT.

de Jaegher, Kris (2008). "The Evolution of Horn's Rule". In: *Journal of Economic Methodology* 15.3. Pp. 275–284.

de Jaegher, Kris et al. (2008). "Economic Laboratory Experiment on Horn's Rule". Unpublished manuscript, Utrecht School of Economics.

Jäger, Gerhard (2002). "Some Notes on the Formal Properties of Bidirectional Optimality Theory". In: *Journal of Logic, Language and Information* 11.4. Pp. 427–451.

— (2007). "Game Dynamics Connects Semantics and Pragmatics". In: *Game Theory and Linguistic Meaning*. Ed. by Ahti-Veikko Pietarinen. Elsevier. Pp. 89–102.

— (2008*a*). "Applications of Game Theory in Linguistics". In: *Language and Linguistics Compass* 2/3. Pp. 406–421.

— (2008*b*). "Game-Theoretical Pragmatics". Unpublished manuscript, University of Bielefeld.

— (2008*c*). "Game Theory in Semantics and Pragmatics". Unpublished manuscript, University of Bielefeld.

Jäger, Gerhard and Christian Ebert (2009). "Pragmatic Rationalizability". In: *Proceedings of Sinn und Bedeutung 13*. Ed. by Arndt Riester and Torgrim Solstad. Pp. 1–15.

de Jager, Tikitu (2009). " 'Now that you mention it I wonder. . . ': Awareness, Attention, Assumption". PhD thesis. Universiteit van Amsterdam.

de Jager, Tikitu and Robert van Rooij (2007). "Explaining Quantity Implicatures". In: *Proceedings of the 11th conference on Theoretical Aspects of Rationality and Knowledge*. New York: ACM. Pp. 193–202.

Johnson-Laird, P. N. (1986). "Conditionals and Mental Models". In: *On Conditionals*. Ed. by Elizabeth Closs Traugott et al. Cambridge University Press. Pp. 55–75.

Johnson-Laird, P. N. et al. (1999). "Naive Probability: A Mental Model Theory of Extensional Reasoning". In: *Psychological Review* 106.1. Pp. 62–88.

Kamp, Hans (1973). "Free Choice Permission". In: *Proceedings of the Aristotelian Society* 74. Pp. 57–74.

— (1978). "Semantics versus Pragmatics". In: *Formal Semantics and Pragmatics for Natural Languages*. Ed. by Franz Guenthner and Siegfried Josef Schmidt. Dordrecht: Reidel. Pp. 255–287.

Karttunen, Lauri and Stanley Peters (1974). "Conventional Implicature". In: *Syntax and Semantics*. Ed. by Choon-Kyu Oh and David A. Dinneen. Vol. 11: Presupposition. Academic Press. Pp. 1–56.

Kasher, Asa (1976). "Conversational Maxims and Rationality". In: *Language in Focus: Foundations, Methods and Systems*. Ed. by Asa Kasher. Dordrech: Reidel. Pp. 197–216.

Katsos, Napoleon (2008*a*). "Evaluating Under-Informative Utterances with Context-Dependent and Context-Independent Scales: Experimental and

Theoretical Implications". To appear in 'Experimental Semantics and Pragmatics.' Ed. by U. Sauerland & K. Yatshushiro.

— (2008*b*). "The Semantics/Pragmatics Interface from an Experimental Perspective: The Case of Scalar Implicature". In: *Synthese* 165.3. Pp. 385–401.

Katsos, Napoleon and Dorothy V. M. Bishop (2009). "The Development of Informativeness from a Speaker's and a Comprehender's Perspective". Unpublished manuscript, University of Cambridge.

Katzir, Roni (2007). "Structurally-Defined Alternatives". In: *Linguistics and Philosophy* 30.6.

Kaufmann, Stefan (2005*a*). "Conditional Predictions — A Probabilistic Account". In: *Linguistics and Philosophy* 28. Pp. 181–231.

— (2005*b*). "Conditionals". In: *Encyclopedia of Language and Linguistics*. Ed. by Keith Brown. Vol. 3. Elsevier. Pp. 6–9.

Keynes, John Maynard (1921). *A Treatise on Probability*. Macmillan.

— (1936). *The General Theory of Employment, Interest, and Money*. London: Macmillan.

Keysar, Boaz et al. (2003). "Limits on Theory of Mind Use in Adults". In: *Cognition* 89.1. Pp. 25–41.

Klinedinst, Nathan (2006). "Plurality and Possibility". PhD thesis. University of California, Los Angeles.

Kohlberg, Elon and Jean-François Mertens (1986). "On the Strategic Stability of Equilibria". In: *Econometrica* 54.5. Pp. 1003–1037.

Köpcke, Klaus-Michael and Klaus-Uwe Panther (1989). "On Correlations between Word Order and Pragmatic Function of Conditional Sentences in German". In: *Journal of Pragmatics* 13. Pp. 685–711.

Kratzer, Angelika (1981). "The Notional Category of Modality". In: *Words, Worlds, and Contexts: New Approaches in Word Semantics*. Ed. by H. J. Eikmeyer and H. Rieser. Berlin: de Gruyter. Pp. 38–74.

— (1991). "Modality". In: *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*. Ed. by Arnim von Stechow and Dieter Wunderlich. Berlin: Walter de Gruyter. Pp. 639–650.

Kratzer, Angelika and Junko Shimoyama (2002). "Indeterminate Pronouns: The View from Japanese". In: *Proceeding of the 3rd Tokyo Conference on Psycholinguistics*. Ed. by Yukio Otsu. Pp. 1–25.

Kreps, David M. and Robert Wilson (1982). "Sequential Equilibria". In: *Econometrica* 50.4. Pp. 863–894.

van Kuppevelt, Jan (1996). "Inferring from Topics: Scalar Implicatures as Topic-Dependent Inferences". In: *Linguistics and Philosophy* 19.4. Pp. 393–443.

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

— (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicature*. Cambridge, Massachusetts: MIT Press.

Lewis, David (1969). *Convention. A Philosophical Study*. Harvard University Press.

— (1973). *Counterfactuals*. Harvard University Press.

— (1975). "Adverbs of Quantification". In: *Formal Semantics of Natural Language*. Ed. by Edward L. Keenan. Cambridge University Press. Pp. 178–188.

— (1981). "Ordering Semantics and Premise Semantics for Counterfactuals". In: *Journal of Philosophical Logic* 10.2. Pp. 217–234.

— (1988). "Relevant Implication". In: *Theoria* 54. Pp. 162–174.

Matsui, Akihiko (1992). "Best Response Dynamics and Socially Stable Strategies". In: *Journal of Economic Theory* 57.2. Pp. 343–362.

Matsumoto, Yo (1995). "The Conversational Condition on Horn Scales". In: *Linguistics and Philosophy* 18.1. Pp. 21–60.

Matthews, Steven A. et al. (1991). "Refining Cheap Talk Equilibria". In: *Journal of Economic Theory* 55. Pp. 247–273.

McCawley, James D. (1981). *Everything that Linguists Have Always Wanted to Know About Logic but Were Ashamed to Ask*. University of Chicago Press.

— (1996). "Conversational Scorekeeping and the Interpretation of Conditionals". In: *Grammatical Constructions*. Ed. by Masayoshi Shibatani and Sandra A. Thompson. New York: Oxford University Press. Pp. 77–101.

McClure, William (2000). *Using Japanese — A Guide to Contemporary Usage*. Cambridge University Press.

McKay, Thomas and Peter van Inwagen (1977). "Counterfactuals with Disjunctive Antecedents". In: *Philosophical Studies* 31.5. Pp. 353–356.

Merin, Arthur (1992). "Permission Sentences Stand in the Way of Boolean and Other Lattice-Theoretic Semantices". In: *Journal of Semantics* 9. Pp. 95–162.

— (1999). "Information, Relevance, and Social Decisionmaking: Some Principles and Results of Decision-Theoretic Semantics". In: *Logic, Language and Computation*. Ed. by Lawrence C. Moss et al. Vol. 2. CSLI Publications. Pp. 179–221.

Mill, John Stuart (1867). *An Examination of Sir William Hamilton's Philosophy*. 3rd ed. London: Longman.

Milne, Alan Alexander (1991). *The Complete Winnie-the-Pooh*. London: Dean.
Morreau, Michael (1997). "Fainthearted Conditionals". In: *Journal of Philosophy* 94.4. Pp. 187–211.
Morris, Charles M. (1946). *Signs, Language, and Behavior*. New York: Prentice Hall.
Myerson, Roger B. (1989). "Credible Negotiation Statements and Coherent Plans". In: *Journal of Economic Theory* 48.1. Pp. 264–303.
— (1991). *Game Theory: Analysis of Conflict*. Harvard University Press.
Nagel, Rosemarie (1995). "Unravelling in Guessing Games: An Experimental Study". In: *The American Economic Review* 85.5. Pp. 1313–1326.
Neale, Stephen (1992). "Paul Grice and the Philosophy of Language". In: *Linguistics and Philosophy* 15. Pp. 509–559.
Newstead, Stephen E. et al. (1997). "Conditional Reasoning with Realistic Material". In: *Thinking and Reasoning* 3.1. Pp. 49–76.
Noh, Eun-Ju (1998). "A Relevance-Theoretic Account of Metarepresentatitive Uses in Conditionals". In: *Current Issues in Relevance Theory*. Ed. by Villy Rouchota and Andreas H. Jucker. John Benjamins. Pp. 271–304.
Noveck, Ira A. (2001). "When Children are more Logical than Adults: Experimental Investigations of Scalar Implicature". In: *Cognition* 78. Pp. 165–188.
Noveck, Ira A. and Andres Posada (2003). "Characterizing the Time Course of an Implicature: An Evoked Potentials Study". In: *Brain and Language* 85.2. Pp. 203–210.
Noveck, Ira A. and Dan Sperber, eds. (2004). *Experimental Pragmatics*. Hampshire: Palgrave MacMillan.
Nute, Donald (1975). "Counterfactuals and the Similarity of Worlds". In: *Journal of Philosophy* 21.4. Pp. 773–778.
O'Hair, S. G. (1969). "Implications and Meaning". In: *Theoria* 35.1. Pp. 38–54.
Osborne, Martin J. (2004). *An Introduction to Game Theory*. New York: Oxford University Press.
Osborne, Martin J. and Ariel Rubinstein (1994). *A Course in Game Theory*. MIT Press.
Papafragou, Anna and Julien Musolino (2003). "Scalar Implicatures: Experiments at the Semantics-Pragmatics Interface". In: *Cognition* 86. Pp. 253–282.
Parikh, Prashant (1991). "Communication and Strategic Inference". In: *Linguistics and Philosophy* 473–514.14. P. 3.
— (1992). "A Game-Theoretic Account of Implicature". In: *TARK '92: Proceedings of the 4th conference on Theoretical aspects of reasoning about knowledge*. San Francisco: Morgan Kaufmann Publishers Inc. Pp. 85–94.

Parikh, Prashant (2001). *The Use of Language*. Stanford University: CSLI Publications.

— (2006). "Pragmatics and Games of Partial Information". In: *Game Theory and Pragmatics*. Ed. by Anton Benz et al. Palgrave MacMillan. Pp. 101–122.

Pauw, Simon (2008). "A BiOT Account of Gricean Reasoning". MA thesis. Universiteit van Amsterdam.

Pearce, David G. (1984). "Rationalizable Strategic Behavior and the Problem of Perfection". In: *Econometrica* 52.4. Pp. 1029–1050.

Poesio, Massimo (1996). "Semantic Ambiguity and Perceived Ambiguity". In: *Semantic Ambiguity and Underspecification*. Ed. by Kees van Deemter and Stanley Peters. CSLI Publications. Pp. 159–201.

Potts, Chris (2005). *The Logic of Conventional Implicatures*. Oxford University Press.

Pouscoulous, Nausicaa et al. (2007). "A Developmental Investigation of Processing Costs in Implicature Production". In: *Language Acquisition* 14.4. Pp. 347–375.

Predelli, Stefano (2007). "Towards a Semantics for Biscuit Conditionals". In: *Philosophical Studies* 142.3. Pp. 293–305.

Premack, David and Guy Woodruff (1978). "Does the Chimpanzee have a Theory of Mind". In: *Behavioral and Brain Sciences* 1.4. Pp. 515–526.

Prince, Alan and Paul Smolensky (1997). "Optimality: From Neural Networks to Universal Grammar". In: *Science* 275. Pp. 1604–1610.

Rabin, Matthew (1990). "Communication between Rational Agents". In: *Journal of Economic Theory* 51. Pp. 144–170.

— (1992). "Corrigendum". In: *Journal of Economic Theory* 58. Pp. 110–111.

— (1994). "A Model of Pre-Game Communication". In: *Journal of Economic Theory* 63.2. Pp. 370–391.

Ramsey, Frank Plumpton (1931). "General Propositions and Causality". In: *The Foundations of Mathematics and other Logical Essays*. Ed. by R. B. Braithwaite. London: Kegan Paul, Trench and Trubner. Pp. 237–255.

Rawlins, Kyle (2008*a*). "(Un)conditionals: An Investigation in the Syntax and Semantics of Conditional Structures". PhD thesis. University of California, Santa Cruz.

— (2008*b*). "Unifying *If*-Conditinals and Unconditionals". To appear in Proceedings of SALT XVIII.

Recanati, François (1989). "The Pragmatics of What is Said". In: *Mind and Language* 4.4. Pp. 295–239.

— (2004). *Literal Meaning*. Cambridge: Cambridge University Press.

Roberts, Craige (1989). "Modal Subordination and Pronominal Anaphora in Discourse". In: *Linguistics and Philosophy* 12. Pp. 683–721.

van Rooij, Robert (2000). "Permission to Change". In: *Journal of Semantics* 17.2. Pp. 119–143.

— (2003*a*). "Quality and Quantity of Information Exchange". In: *Journal of Logic, Language and Computation* 12. Pp. 423–451.

— (2003*b*). "Questioning to Resolve Decision Problems". In: *Linguistics and Philosophy* 29. Pp. 727–763.

— (2004*a*). "Cooperative versus Argumentative Communication". In: *Philosophia Scientiae* 8.2. Pp. 195–209.

— (2004*b*). "Signalling Games Select Horn-Strategies". In: *Linguistics and Philosophy* 27. Pp. 493–527.

— (2004*c*). "Utility, Informativity and Protocols". In: *Journal of Philosophical Logic* 33.4. Pp. 389–419.

— (2006*a*). "Free Choice Counterfactual Donkeys". In: *Journal of Semantics* 23.4. Pp. 383–402.

— (2006*b*). *Optimality-Theoretic and Game-Theoretic Approaches to Implicatures*. Stanford Encyclopedia of Philosophy.

— (2007). "Strengthening Conditional Presuppositions". In: *Journal of Semantics* 24.3. Pp. 289–304.

— (2008). "Games and Quantity Implicatures". In: *Journal of Economic Methodology* 15.3. Pp. 261–274.

van Rooij, Robert and Katrin Schulz (2004). "Exhaustive Interpretation of Complex Sentences". In: *Journal of Logic, Language and Information* 13. Pp. 491–519.

— (2006). "Only: Meaning and Implicatures". In: *Questions in Dynamic Semantics*. Ed. by Maria Aloni et al. Amsterdam, Singapore: Elsevier. Pp. 193–223.

Rothschild, Daniel (2008). "Grice, Utterance Choice, and Rationality". Unpublished manuscript, Columbia University.

Rubinstein, Ariel (1989). "The Electronic Mail Game: Strategic Behavior Under 'Almost Common Knowledge' ". In: *The American Economic Review* 79.3. Pp. 385–391.

Rubinstein, Ariel et al. (1996). "Naïve Strategies in Competitive Games". In: *Understanding Strategic Interaction — Essays in Honor of Reinhard Selten*. Ed. by Wulf Albers et al. Berlin: Springer Verlag. Pp. 394–102.

Russell, Benjamin (2006). "Against Grammatical Computation of Scalar Implicatures". In: *Journal of Semantics* 23.361–382.

Rutherford, William (1970). "Some observations concerning subordinate clauses in English". In: *Language* 46. Pp. 97–115.

Sally, David (2003). "Risky Speech: Behavioral Game Theory and Pragmatics". In: *Journal of Pragmatics* 35. Pp. 1223–1245.

Sampson, Geoffrey (1982). "The Economics of Conversation". In: *Mutual Knowledge*. Ed. by Neilson Voyne Smith. Academic Press. Pp. 200–210.

Sauerland, Uli (2004). "Scalar Implicatures in Complex Sentences". In: *Linguistics and Philosophy* 27. Pp. 367–391.

Savant, Marilyn Vos (1994). *Ask Marilyn*. St Martins Mass Market Paper.

Scheffler, Tatjana (2008*a*). "Relevance Conditionals as Utterance Modifying Adverbials". In: *Empirical Issues in Syntax and Semantics 7*. Ed. by Olivier Bonami and Patricia Cabredo Hofherr. Pp. 373 –392.

— (2008*b*). "Semantic Operators in Different Dimensions". PhD thesis. University of Pennsylvania.

Schelling, Thomas C. (1960). *The Strategy of Conflict*. Cambridge, Massachusetts: Harvard University Press.

Schulz, Katrin (2005). "A Pragmatic Solution for the Paradox of Free Choice Permission". In: *Synthese* 147. Pp. 343–377.

— (2007). "Minimal Models in Semantics and Pragmatics — Free Choice, Exhaustivity, and Conditionals". PhD thesis. Universiteit van Amsterdam.

Schulz, Katrin and Robert van Rooij (2006). "Pragmatic Meaning and Nonmonotonic Reasoning: The Case of Exhaustive Interpretation". In: *Linguistics and Philosophy* 29. Pp. 205–250.

Selten, Reinhard (1965). "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit". In: *Zeitschrift für die die gesamte Staatswissenschaft* 121. Pp. 301–324.

— (1975). "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games". In: *International Journal of Game Theory* 4. Pp. 25–55.

— (1998). "Features of Experimentally Observed Bounded Rationality". In: *European Economic Review* 42. Pp. 413–436.

Shimoji, Makoto (2002). "On Forward Induction in Money Burning Games". In: *Economic Theory* 19. Pp. 637–648.

Siegel, Muffy E. A. (2006). "Biscuit Conditionals: Quantification over potential literal acts". In: *Linguistics and Philosophy* 29. Pp. 167–203.

Simons, Mandy (2005). "Dividing Things Up: The Semantics of *or* and the modal/*or* interaction". In: *Natural Language Semantics* 13.3. Pp. 271–316.

Smith, Carol L. (1980). "Quantifiers and Question Answering in Young Children". In: *Journal of Experimental Child Psychology* 30.2. Pp. 191–205.

Smolensky, Paul (1996). "On the Comprehension/Production Dilemma in Child Language". In: *Linguistic Inquiry* 27.4. Pp. 720–731.

Soames, Scott (1982). "How Presuppositions are Inherited: A Solution to the Projection Problem". In: *Linguistic Inquiry* 13.3. Pp. 483–545.

Sobel, Joel (1985). "A Theory of Credibility". In: *Review of Economic Studies* 52.4. Pp. 557–573.

— (to appear). "Signaling Games". In: *Encyclopedia of Complexity and Systems Science*. Ed. by M. Sotomayor. Springer Verlag.

Spector, Benjamin (2006). "Scalar Implicatures: Exhaustivity and Gricean Reasoning". In: *Questions in Dynamic Semantics*. Ed. by Maria Aloni et al. Amsterdam, Singapore: Elsevier. Pp. 229–254.

Spence, Andrew Michael (1973). "Job market signaling". In: *Quarterly Journal of Economics* 87. Pp. 355–374.

Sperber, Dan and Deirde Wilson (1995). *Relevance: Communication and Cognition (2nd ed.)* Oxford: Blackwell.

— (2004). "Relevance Theory". In: *Handbook of Pragmatics*. Ed. by Laurence R. Horn and Gregory Ward. Oxford: Blackwell. Pp. 607–632.

Stahl, Dale O. (1993). "Evolution of Smart$_n$ Players". In: *Games and Economic Behavior* 5.4. Pp. 604–617.

Stahl, Dale O. and Paul W. Wilson (1995). "On Players' Models of Other Players: Theory and Experimental Evidence". In: *Games and Economic Behavior* 10. Pp. 218–254.

Stalnaker, Robert (1968). "A Theory of Conditionals". In: *Studies in Logical Theory*. Ed. by Nicholas Rescher. Vol. 2. Oxford University Press. Pp. 98–112.

— (1994). "On the Evaluation of Solution Concepts". In: *Theory and Decision* 37. Pp. 49–73.

— (1998). "Belief Revision in Games: Forward and Backward Induction". In: *Mathematical Social Sciences* 36. Pp. 31–56.

— (2006). "Saying and Meaning, Cheap Talk and Credibility". In: *Game Theory and Pragmatics*. Ed. by Anton Benz et al. Hampshire: Palgrave MacMillan. Pp. 83–100.

Swanson, Eric (2003). "Biscuit Conditionals and the Common Ground". In: *Proceedings of the Student Session of NASSLLI*. Ed. by John Hale. Pp. 26–34.

Sweetser, Eve (1990). *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press.

Tversky, Amos and Daniel Kahnemann (1983). "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment". In: *Psychological Review* 90.4. Pp. 293–315.

Veltman, Frank (1985). "Logics for Conditionals". PhD thesis. Universiteit van Amsterdam.

— (1986). "Data Semantics and the Pragmatics of Conditionals". In: *On Conditionals*. Ed. by Elizabeth Closs Traugott et al. Cambridge University Press. Pp. 147–168.

— (2005). "Making Counterfactual Assumptions". In: *Journal of Semantics* 22.2. Pp. 159–180.

Walker, Willard (1970). "The Retention of Folk Linguistic Concepts and the *ti'yčir* Caste in Contemporary Nacireman Culture". In: *American Anthropologist* 72.1. Pp. 102–105.

Warmbrōd, Ken (1981). "Counterfactuals and Substitution of Equivalent Antecedents". In: *Journal of Philosophical Logic* 10.2. Pp. 267–289.

Weibull, Jörgen W. (1997). *Evolutionary Game Theory*. MIT Press.

Wilson, Deirde and Tomoko Matsui (1998). "Recent Approaches to Bridging: Truth, Coherence and Relevance". In: *UCL Working Papers in Linguistics*. Vol. 10. Pp. 173–200.

Wimmer, Heinz and Josef Perner (1983). "Beliefs about Beliefs: Representing and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception". In: *Cognition* 13.1. Pp. 103–128.

Zahavi, Amotz (1975). "Mate Selection — A Selection for a Handicap". In: *Journal of Theoretical Biology* 53. Pp. 205–214.

Zapater, Inigo (1997). "Credible Proposals in Communication Games". In: *Journal of Economic Theory* 72. Pp. 173–197.

Zeevat, Henk (2000). "The Asymmetry of Optimality Theoretic Syntax and Semantics". In: *Journal of Semantics* 17.3. Pp. 243–262.

Zhang, Jun and Trey Hedden (2003). "Two Paradigms for Depth of Strategic Reasoning in Games". In: *Trends in Cognitive Sciences* 7.1. Pp. 4–5.

Zimmermann, Thomas Ede (2000). "Free Choice Disjunction and Epistemic Possibility". In: *Natural Language Semantics* 8. Pp. 255–290.

Zondervan, Arjen (2006). "The Question Under Discussion Focus Condition for Scalar Implicatures". MA thesis. Universiteit Utrecht.

Zwarts, Joost (2006). "*Om* en *rond*: Een Semantische Vergelijking". In: *Nederlandse Taalkunde* 11.2. Pp. 101–123.

Notation, Symbols & Abbreviations

$\mathbb{N}$: natural numbers

$\mathbb{R}$: real numbers

$\rightsquigarrow$: standard implicature (p. 6)

$\Delta(X)$: set of probability distributions on set X (p. 19)

X^Y: set of all functions from Y to X (p. 19)

$X : Y \to Z$: alternative for $X \in Z^Y$ (p. 19)

$\mathcal{P}(X)$: power set of set X (p. 19)

$[\![\cdot]\!]$: semantic denotation (p. 19)

S, R: sender, receiver (p. 19)

T: set of states (p. 19)

M: set of messages (p. 19)

A: set of actions (p. 19)

$\mathrm{U}_{S,R}$: utility functions (p. 19)

$\mathrm{V}_{S,R}$: response utilities (p. 19)

$\mathrm{C}_{S,R}$: message costs (p. 19)

S, R: sets of all pure strategies (p. 24)

s, r: pure strategy (p. 24)

$\mathcal{S}$, $\mathcal{R}$: sets of all probabilistic strategies (p. 24)

σ, ρ: probabilistic strategy (p. 24)
or behavioral belief (p. 26)

μ: posterior receiver belief (p. 26)

$\mathrm{EU}_{S,R}$: expected utility (p. 28)

$\mathrm{BR}(\cdot)$: best response to a belief (p. 29)

$\mathrm{Rat}_{S,R}$: all rationalizable strategies (p. 30)

$A^*(m)$: zero-order rationalizable actions (p. 56)

S_k, R_k: sender/receiver of level k (p. 54)

S^*, R^*: limit prediction of the IBR model (p. 58)

$A > C$: conditional "if A, then C" (p. 227)

GTP: game theoretic pragmatics (p. 12)

PBE: perfect Bayesian equilibrium (p. 31)

IBR: iterated best response

TCP: truth ceteris paribus (p. 72)

FI: forward induction (p. 83)

BR: best response (p. 98)

CURB: closed under rational behavior (p. 99)

CMR: credible message rationalizability (p. 118)

FC: free choice (p. 156)

SDA: simplification of disjunctive antecedents (p. 169)

OT: optimality theory (p. 182)

BIOT: bidirectional optimality theory (p. 182)

CP: conditional perfection (p. 234)

BC: biscuit conditional (p. 257)

CBC: counterfactual biscuit conditional (p. 270)

Samenvatting

Dit proefschrift levert een speltheoretisch model van taalgebruik en interpretatie en past het toe op Griceaanse pragmatiek (Grice 1989). Het model dat hier gepresenteerd wordt, het zogenoemde IBR model, verklaart pragmatische verschijnselen, zoals conversationele implicaturen, als resultaat van een sequentie van herhaalde optimaliseringsstappen (Engels: iterated best responses). Het model beschouwt de letterlijke, semantische betekenis van uitdrukkingen als een centraal element in het denkpatroon van spreker en hoorder. Zo wordt aanvankelijk het gedrag van spelers in een signaalspel bepaald door alleen of voornamelijk semantische informatie in aanmerking te nemen. Vervolgens kunnen taalgebruikers dan hun gedrag aanpassen aan zo een verondersteld letterlijk taalgebruik, om hun conversationele doelen optimaal te kunnen realiseren. Als deze stap herhaald wordt, ontsaat een proces van optimaliseren gebaseerd op het geoptimaliseerde gedrag van anderen.

Formeel gezien, mag het IBR model beschouwd worden als een versie van *strong rationalizability* in signaalspelen (Battigalli 2006). De aannames over de psychologie van spelers die het IBR model toevoegt zijn gebaseerd op recent empirisch onderzoek (zie Stahl and Wilson 1995; Ho et al. 1998; Camerer et al. 2004), en leiden tot een op natuurlijke wijze beperkt, nieuw en simpel oplossingsconcept dat gerelateerde speltheoretische benaderingen samenvat en verder ontwikkelt (vergelijk in het bijzonder Benz 2006; Stalnaker 2006; Benz and van Rooij 2007; Jäger 2007). Op die manier kunnen relevante verschijnsels, zoals scalare implicaturen, M-implicaturen, en ook zogenoemde *free choice* lezingen, verklaart worden. Bovendien is het IBR model ook geschikt om verschillende vormen van sub-optimaal taalgebruik weer te geven. Dit is belangrijk om bijvoorbeeld te kunnen verklaren hoe zich pragmatische taal-

vaardigheden ontwikkelen binnen de taalverwerving.

Abstract

This thesis offers a general game theoretic model of language use and interpretation and applies it to linguistic pragmatics in the vein of Grice (1989). The model presented here —called the IBR model— explains pragmatic phenomena, such as conversational implicatures, as arising from a sequence of iterated best responses: starting from the literal, semantic meaning as a psychologically salient attractor of attention, speaker and hearer initially compute the rational best responses to a literal use or interpretation of expressions; subsequently, agents continue computing best responses to best responses, for as long as this is reasonable and their cognitive resources permit.

This algorithmic solution procedure is simple and intuitively appealing. But more importantly, it has a clear epistemic interpretation as modelling so-called "level-k thinking," which has gained recent popularity in behavioral game theory (Stahl and Wilson 1995; Ho et al. 1998; Camerer et al. 2004). Laboratory data supports the assumption that human reasoners are cognitively biased and possibly resource-bounded in the sense that they are susceptible to focal framing effects and perform theory of mind reasoning possibly only to a given depth k. Thus conceived, the IBR model formally implements a number of empirically attested assumptions about the cognitive architecture of human reasoners. The IBR model then effectively provides a novel non-equilibrium solution concept as a form of strong rationalizability (Battigalli 2006) in which these psychological assumptions have been implemented. The thesis aims to show how this turn towards psychological realism solves outstanding conceptual problems with game theoretic approaches to communication, and, moreover, improves on predictions in linguistic applications.

Firstly, by implementing semantic meaning as a focal attractor of atten-

tion, the IBR model singles out those strategies that conform to our intuitions about credible communication without altogether precluding the possibility, and even occasional optimality, of lying, misleading and distrust (see Farrell and Rabin 1996; Stalnaker 2006). Secondly, the model explicitly represents agents with absent or only limited capacity of taking opponent behavior and reasoning into account. This sheds light on higher-order theory of mind reasoning in language use and especially in the pattern of acquisition of pragmatic competence by young children (see Noveck 2001; Papafragou and Musolino 2003). An in-depth comparison of the IBR model with bidirectional optimality theory (Blutner 2000) suggests that the former is the better tool for modelling limitations of theory of mind reasoning in interpretation and acquisition. Finally, the IBR model unifies and extends a series of recent work in game theoretic pragmatics (see especially Benz 2006; Stalnaker 2006; Benz and van Rooij 2007; Jäger 2007). It yields formidable predictions for, among other phenomena, complex and nested cases of scalar implicatures, generalized M-implicatures and free-choice readings. The model also backs up natural accounts of conditional perfection, and unconditional readings of conditionals.

Titles in the ILLC Dissertation Series:

ILLC DS-2001-01: **Maria Aloni**
Quantification under Conceptual Covers

ILLC DS-2001-02: **Alexander van den Bosch**
Rationality in Discovery - a study of Logic, Cognition, Computation and Neuropharmacology

ILLC DS-2001-03: **Erik de Haas**
Logics For OO Information Systems: a Semantic Study of Object Orientation from a Categorial Substructural Perspective

ILLC DS-2001-04: **Rosalie Iemhoff**
Provability Logic and Admissible Rules

ILLC DS-2001-05: **Eva Hoogland**
Definability and Interpolation: Model-theoretic investigations

ILLC DS-2001-06: **Ronald de Wolf**
Quantum Computing and Communication Complexity

ILLC DS-2001-07: **Katsumi Sasaki**
Logics and Provability

ILLC DS-2001-08: **Allard Tamminga**
Belief Dynamics. (Epistemo)logical Investigations

ILLC DS-2001-09: **Gwen Kerdiles**
Saying It with Pictures: a Logical Landscape of Conceptual Graphs

ILLC DS-2001-10: **Marc Pauly**
Logic for Social Software

ILLC DS-2002-01: **Nikos Massios**
Decision-Theoretic Robotic Surveillance

ILLC DS-2002-02: **Marco Aiello**
Spatial Reasoning: Theory and Practice

ILLC DS-2002-03: **Yuri Engelhardt**
The Language of Graphics

ILLC DS-2002-04: **Willem Klaas van Dam**
On Quantum Computation Theory

ILLC DS-2002-05: **Rosella Gennari**
Mapping Inferences: Constraint Propagation and Diamond Satisfaction

ILLC DS-2002-06: **Ivar Vermeulen**
A Logical Approach to Competition in Industries

ILLC DS-2003-01: **Barteld Kooi**
Knowledge, chance, and change

ILLC DS-2003-02: **Elisabeth Catherine Brouwer**
Imagining Metaphors: Cognitive Representation in Interpretation and Understanding

ILLC DS-2003-03: **Juan Heguiabehere**
Building Logic Toolboxes

ILLC DS-2003-04: **Christof Monz**
From Document Retrieval to Question Answering

ILLC DS-2004-01: **Hein Philipp Röhrig**
Quantum Query Complexity and Distributed Computing

ILLC DS-2004-02: **Sebastian Brand**
Rule-based Constraint Propagation: Theory and Applications

ILLC DS-2004-03: **Boudewijn de Bruin**
Explaining Games. On the Logic of Game Theoretic Explanations

ILLC DS-2005-01: **Balder David ten Cate**
Model theory for extended modal languages

ILLC DS-2005-02: **Willem-Jan van Hoeve**
Operations Research Techniques in Constraint Programming

ILLC DS-2005-03: **Rosja Mastop**
What can you do? Imperative mood in Semantic Theory

ILLC DS-2005-04: **Anna Pilatova**
A User's Guide to Proper names: Their Pragmatics and Semanics

ILLC DS-2005-05: **Sieuwert van Otterloo**
A Strategic Analysis of Multi-agent Protocols

ILLC DS-2006-01: **Troy Lee**
Kolmogorov complexity and formula size lower bounds

ILLC DS-2006-02: **Nick Bezhanishvili**
Lattices of intermediate and cylindric modal logics

ILLC DS-2006-03: **Clemens Kupke**
Finitary coalgebraic logics

ILLC DS-2006-04: **Robert Špalek**
Quantum Algorithms, Lower Bounds, and Time-Space Tradeoffs

ILLC DS-2006-05: **Aline Honingh**
The Origin and Well-Formedness of Tonal Pitch Structures

ILLC DS-2006-06: **Merlijn Sevenster**
Branches of imperfect information: logic, games, and computation

ILLC DS-2006-07: **Marie Nilsenova**
Rises and Falls. Studies in the Semantics and Pragmatics of Intonation

ILLC DS-2006-08: **Darko Sarenac**
Products of Topological Modal Logics

ILLC DS-2007-01: **Rudi Cilibrasi**
Statistical Inference Through Data Compression

ILLC DS-2007-02: **Neta Spiro**
What contributes to the perception of musical phrases in western classical music?

ILLC DS-2007-03: **Darrin Hindsill**
It's a Process and an Event: Perspectives in Event Semantics

ILLC DS-2007-04: **Katrin Schulz**
Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals

ILLC DS-2007-05: **Yoav Seginer**
Learning Syntactic Structure

ILLC DS-2008-01: **Stephanie Wehner**
Cryptography in a Quantum World

ILLC DS-2008-02: **Fenrong Liu**
Changing for the Better: Preference Dynamics and Agent Diversity

ILLC DS-2008-03: **Olivier Roy**
Thinking before Acting: Intentions, Logic, Rational Choice

ILLC DS-2008-04: **Patrick Girard**
Modal Logic for Belief and Preference Change

ILLC DS-2008-05: **Erik Rietveld**
Unreflective Action: A Philosophical Contribution to Integrative Neuroscience

ILLC DS-2008-06: **Falk Unger**
Noise in Quantum and Classical Computation and Non-locality

ILLC DS-2008-07: **Steven de Rooij**
Minimum Description Length Model Selection: Problems and Extensions

ILLC DS-2008-08: **Fabrice Nauze**
Modality in Typological Perspective

ILLC DS-2008-09: **Floris Roelofsen**
Anaphora Resolved

ILLC DS-2008-10: **Marian Counihan**
Looking for logic in all the wrong places: an investigation of language, literacy and logic in reasoning

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder…": Awareness, Attention, Assumption