Signal to act : game theory in pragmatics
Franke, M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
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
Chapter 1. What is Game Theoretic Pragmatics?

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).\(^3\)

\[
\begin{align*}
(3) \quad & \text{a. I saw some of your children today.} \\
& \text{b. The speaker did not see all of the hearer’s children today.}
\end{align*}
\]

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).

\[
\begin{align*}
(4) \quad & \text{I saw some of your children today, and maybe even all of them.}
\end{align*}
\]

---

\(^2\) 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):

\[
\begin{align*}
(1) \quad & \text{The speaker knows/believes that she did not see all of the hearer’s children.} \\
(2) \quad & \text{The speaker does not know/believe that she saw all of the hearer’s children.}
\end{align*}
\]

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). 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. \( \leadsto \) 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 \( \leadsto \) 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. $\sim$ It’s not certain/necessary that Yuuki is coming late again.

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

The abstract reasoning pattern behind a scalar inference seems to be the following **naive scalar reasoning**: 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 naive scalar reasoning pattern.

(8)  
\[ \begin{align*}
\text{a.} & \quad \text{Some of Kiki’s friends are metalheads.} \\
\text{b.} & \quad \text{Few of Kiki’s friends are metalheads.} \\
\text{c.} & \quad \text{Most of Kiki’s friends are metalheads.} \\
\text{d.} & \quad \text{All of Kiki’s friends are metalheads.} \\
\text{e.} & \quad \lnot \quad \text{It’s not the case that few of Kiki’s friends are metalheads.} \\
\text{f.} & \quad \leadsto \quad \text{It’s not the case that most of Kiki’s friends are metalheads.} \\
\text{g.} & \quad \leadsto \quad \text{It’s not the case that all of Kiki’s friends are metalheads.}
\end{align*} \]

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)  
\[ \begin{align*}
\text{a.} & \quad \text{Black Bart killed the sheriff.} \\
\text{b.} & \quad \leadsto \quad \text{Black Bart killed the sheriff in a stereotypical way.}
\end{align*} \]

(10)  
\[ \begin{align*}
\text{a.} & \quad \text{Black Bart caused the sheriff to die.} \\
\text{b.} & \quad \leadsto \quad \text{Black Bart killed the sheriff in a non-stereotypical way.}
\end{align*} \]

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 (\( m \leadsto t \)); the second part of the pragmatic inference pattern associates the marked form with the marked state of affairs (\( m' \leadsto t' \)).
1.1. Gricean Pragmatics

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. 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. In the following, I will specifically use this term to refer to the second part of the inference pattern, the association $m' \sim t'$ 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 Quantity as requirements on speaker and hearer economy (see below). This then also explains the label ‘division of pragmatic labor.’

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-beexplained 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):

\[(11)\quad \text{Q-Principle} \quad \text{Say as much as you can (given I)}.\]

\[(12)\quad \text{I-Principle} \quad \text{Say no more than you must (given Q)}.\]

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:

\[(12)\quad \text{John has a very efficient secretary.} \quad \rightarrow \quad \text{John has a very efficient female secretary.}\]

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:
Chapter 1. What is Game Theoretic Pragmatics?

“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 2003a; 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 2004b;
Jäger 2007).\textsuperscript{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 \textit{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.

\section*{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.\textsuperscript{11} A \textit{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 \textit{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 \textit{solution concept} that describes—or, depending on the preferred interpretation, \textit{prescribes}—actual reasoning and/or decision making. An example of a simple game with some of its possible solution concepts will make this distinction clear.

\subsection*{1.2.1 Static Games & Their Solutions}

\textbf{Prisoner’s Dilemma.} A well-known idealized example situation that game theory models is the so-called \textit{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

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

\textsuperscript{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.
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.
1.2. Game Theory for Gricean Pragmatics

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 2004b).\(^\text{12}\)

---

\(^{12}\) 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).
1.2. Game Theory for Gricean Pragmatics

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

\[ \langle \{S, R\}, T, \Pr, \mathbb{L}, A, U_S, U_R \rangle \]

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; \( \mathbb{L} : M \rightarrow \mathcal{P}(T) \setminus \emptyset \) 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 \( U_{S,R} : T \times M \times A \rightarrow \mathbb{R} \) 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 \( U_{S,R} \) as composed of response utilities \( V_{S,R} : T \times A \rightarrow \mathbb{R} \) minus message costs \( C_{S,R} : T \times M \rightarrow \mathbb{R} \), so that:

\[ U_{S,R}(t, m, a) = V_{S,R}(t, a) - C_{S,R}(t, m). \]

This makes it easier to express ideas such as that messages are entirely costless: \( C_{S,R}(t, m) = 0 \) 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 \( t_{\text{beef}} \) Alice prepares beef, and in \( t_{\text{fish}} \) 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 \( Y \), \( X : Y \rightarrow Z \) 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 \emph{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 \( t \in [m] \).
Chapter 1. What is Game Theoretic Pragmatics?

distribution \( \text{Pr} \). 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 \( m_{\text{beef}} \) “I’m preparing beef” or \( m_{\text{fish}} \) “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 (\( a_{\text{red}} \)) or white wine (\( a_{\text{white}} \)). 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, \( m_{\text{beef}} \) in state \( t_{\text{fish}} \). 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

<table>
<thead>
<tr>
<th>( \text{Pr}(t) )</th>
<th>( a_{\text{red}} )</th>
<th>( a_{\text{white}} )</th>
<th>( m_{\text{beef}} )</th>
<th>( m_{\text{fish}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_{\text{beef}} )</td>
<td>( \frac{3}{5} )</td>
<td>( 1,1 )</td>
<td>( 0,0 )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>( t_{\text{fish}} )</td>
<td>( \frac{2}{5} )</td>
<td>( 0,0 )</td>
<td>( 1,1 )</td>
<td>( - )</td>
</tr>
</tbody>
</table>

Figure 1.2: Signaling game for coordination
1.2. Game Theory for Gricean Pragmatics

<table>
<thead>
<tr>
<th>Pr(t)</th>
<th>a_{\exists \neg \forall}</th>
<th>a_{\forall}</th>
<th>m_{\text{some}}</th>
<th>m_{\text{all}}</th>
</tr>
</thead>
<tbody>
<tr>
<td>t_{\exists \neg \forall}</td>
<td>1 - p</td>
<td>1,1</td>
<td>0,0</td>
<td>√</td>
</tr>
<tr>
<td>t_{\forall}</td>
<td>p</td>
<td>0,0</td>
<td>1,1</td>
<td>√</td>
</tr>
</tbody>
</table>

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.\footnote{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.} The signaling game in figure 1.3 has two states t_{\exists \neg \forall} and t_{\forall}, two messages m_{\text{some}} and m_{\text{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...}
Chapter 1. What is Game Theoretic Pragmatics?

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 \( m_{\text{some}} \) in state \( t_{\exists \forall} \) and the message \( m_{\text{all}} \) in state \( t_{\forall} \); and (ii) the receiver responds to \( m_{\text{some}} \) with \( a_{\text{some}} \) and to \( m_{\text{all}} \) with \( a_{\text{all}} \). 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
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-
Chapter 1. What is Game Theoretic Pragmatics?

Different 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 \( \bar{s} \in \Delta(M^T) \); and a mixed receiver strategy is a probability distribution \( \bar{r} \in \Delta(A^M) \). 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.\footnote{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 $\sigma$ is:}

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

Given a behavioral sender strategy $\sigma$ 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.

**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-
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:

\[ \Pi_S = R = (\Delta(A))^M. \]

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 \( \text{Pr} \) 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 \( \Pi_R \) is the set of all triples \( \langle \text{Pr}, \sigma, \mu \rangle \) 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 $\mu$ 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 $\mu$ to be derived from $Pr$ and $\sigma$ by Bayesian conditionalization. We say that the receiver’s posterior beliefs $\mu$ are consistent with his beliefs $Pr$ and $\sigma$ if and only if for all $t$ in $T$ and for all $m$ in the image of $\sigma$, 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 $\sigma$, i.e., to messages that are expected to be sent under the belief $\sigma$. 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 $\sigma$ 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 \( U : T \times A \to \mathbb{R} \). 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

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

This helps define Bayesian rationality as follows:

\[
(13) \quad \text{Bayesian Rationality} \quad \text{Given an agent’s behavioral alternatives } A, \text{ his beliefs } \Pr \text{ and preferences } U, \text{ the agent is rational only if he chooses an action } a \in A \text{ which maximizes his expected utility (as given by } \Pr \text{ and } U). \]

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 \( \rho \) 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 \( \rho \) as follows:

\[
EU_S(m, t, \rho) = \sum_{a \in A} \rho(m, a) \times U^S(t, m, a).
\]

In line with Bayesian rationality, if \( S \) is rational and believes \( \rho \) she should send a message \( m \) in state \( t \) only if it maximizes her expected utility given belief \( \rho \). We say that a pure sender strategy \( s \in S \) 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 \textbf{RATIONAL PURE SENDER STRATEGY} given belief $\rho$ if and only if for all $t$:

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

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

The receiver’s beliefs are triples $\langle \text{Pr}, \sigma, \mu \rangle \in \Pi_R$, but the important component for a characterization of rational receiver behavior is, of course, the posterior beliefs $\mu$: it’s \textit{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 $\mu$, we define $R$’s expected utility of performing $a$ after message $m$ has been received as

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

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

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

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

\textbf{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.

\textbf{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 $S_0 = S$ and $R_0 = R$, 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:

$$S_{n+1} = \{ s \in S_n \mid \exists \rho \in \Delta(R_n) : s \in BR(\rho) \}$$

$$R_{n+1} = \{ r \in R_n \mid \exists \pi_R = (\Pr, \sigma, \mu) \in \Pi_R :$$

$$\text{(i)} \quad r \in BR(\mu)$$

$$\text{(ii)} \quad \pi_R \text{ is consistent}$$

$$\text{(iii)} \quad \sigma \in \Delta(S_n) \}.$$  

Finally, the sets of rationalizable strategies are the sets

$$\text{Rat}_S = \bigcap_{i \in \mathbb{N}} S_i \quad \text{Rat}_R = \bigcap_{i \in \mathbb{N}} R_i.$$

The set $\text{Rat}_S$ 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 $(s, r)$ 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.\footnote{We will come back to stronger notions of rationalizability in section 2.4.3.} 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
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 (\( \text{pbe} \)) iff three conditions hold:\(^{20}\)

(i) \( \sigma \) is rational given the belief \( \rho \);

(ii) \( \rho \) is rational given the belief \( \mu \);

(iii) \( \mu \) is consistent with \( \Pr \) and the belief \( \sigma \).

In order to check whether a given pure strategy profile \( \langle s, r \rangle \) is a \( \text{pbe} \), we then need to consult the triple \( \langle \sigma, \rho, \mu \rangle \) where \( \sigma \) and \( \rho \) are the unique probabilistic strategies corresponding to the pure strategies \( s \) and \( r \), and where \( \mu \) 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 \( \text{pbe} \). 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 \( \text{pbe} \). 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 \( \text{pbe} \). Let then \( \sigma_1 \) and \( \rho_1 \) 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 \( \pi \) iff, when considered a mixed strategy, \( X \) puts positive probability only on pure best responses to \( \pi \).
Chapter 1. What is Game Theoretic Pragmatics?

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

It is obvious that the only sender strategy which is rational given S’s preferences $U_S$ and the belief $\rho_1$ is $\sigma_1$. Moreover, the receiver’s posterior beliefs are completely determined by the sender’s strategy $\sigma_1$: the only belief $\mu_1$ consistent with any full support prior and the behavioral belief $\sigma_1$ is the posterior belief that puts full credence, i.e., probability 1, on state $t_{\exists\neg\forall}$ after hearing
1.2. Game Theory for Gricean Pragmatics

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

$$
\mu_1 = \left[
\begin{array}{c}
m_{\text{some}} \mapsto \begin{bmatrix} t_{\exists-\forall} & \mapsto & 1 \\ t_{\forall} & \mapsto & 0 \end{bmatrix} \\
m_{\text{all}} \mapsto \begin{bmatrix} t_{\exists-\forall} & \mapsto & 0 \\ t_{\forall} & \mapsto & 1 \end{bmatrix}
\end{array}
\right].
$$

But then, given $\mu_1$ and the receiver’s preferences $U_R$, $\rho_1$ 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 $\sigma_{13}$ 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 $\sigma_6$ and $\rho_6$ 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 $\sigma_6$. Indeed, any belief $\langle \Pr, \sigma_6, \mu_6^q \rangle$ with posterior belief

$$
\mu_6^q = \left[
\begin{array}{c}
m_{\text{some}} \mapsto \begin{bmatrix} t_{\exists-\forall} & \mapsto & \Pr(t_{\exists-\forall}) \\ t_{\forall} & \mapsto & \Pr(t_{\forall}) \end{bmatrix} \\
m_{\text{all}} \mapsto \begin{bmatrix} t_{\exists-\forall} & \mapsto & q \\ t_{\forall} & \mapsto & 1-q \end{bmatrix}
\end{array}
\right]
$$

with $q \in [0,1]$ is consistent. However, not every such posterior belief $\mu_6^q$ makes $\rho_6$ a rational receiver strategy. First of all, $\rho_6$ can only ever be rational if $\Pr(t_{\exists-\forall}) \geq 1/2$. In other words, only for some versions of the some-all game can the profile number 6 be a PBE. Moreover, the receiver strategy $\rho_6$
is rational only for values \( q \geq \frac{1}{2} \). 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 \( \sigma_6 \) and \( \rho_6 \), 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 \( m_{all} \) 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 \( \rho_6 \), which is part of strategy profile 6, is rational only for a posterior belief \( \mu_6 \) for which \( \mu(t_{\exists \neg \forall} | m_{all}) \geq \frac{1}{2} \). But even if \( m_{all} \) 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 \( m_{all} \), no matter whether the receiver expects \( m_{all} \) to be sent or not; hence, any posterior belief that faithfully reflects knowledge of the game structure would set \( \mu(t_{\exists \neg \forall} | m_{all}) = 0 \). 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
Chapter 1. What is Game Theoretic Pragmatics?

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 equilib-rium (Selten 1965, 1975). Another option is sequential equilibrium (Kreps and Wilson 1982) which requires, rather technically, that the receiver’s posterior beliefs $\mu$ 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 $\rho_{16}$ would rather incur her cost in order to coordinate on proper interpretation of the (false) message $m_{a_{11}}$ in state $t_{\exists:\forall}^\downarrow$; if $c = 1$ the sender is indifferent, and so it is still rational to use $m_{a_{11}}$ in state $t_{\exists:\forall}^\downarrow$. 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 $m_{a_{11}}$ in state $m_{\exists:\forall}$ even when $m_{a_{11}}$ is not expected in the first place. 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
Chapter 1. What is Game Theoretic Pragmatics?

<table>
<thead>
<tr>
<th></th>
<th>$a_{\text{heads}}$</th>
<th>$a_{\text{tails}}$</th>
<th>$m_{\text{heads}}$</th>
<th>$m_{\text{tails}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{\text{heads}}$</td>
<td>0, 1</td>
<td>1, 0</td>
<td>$\sqrt{\cdot}$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td>$t_{\text{tails}}$</td>
<td>1, 0</td>
<td>0, 1</td>
<td>$\cdot$</td>
<td>$\sqrt{\cdot}$</td>
</tr>
</tbody>
</table>

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.
1.2. Game Theory for Gricean Pragmatics

<table>
<thead>
<tr>
<th></th>
<th>$a_{heads}$</th>
<th>$a_{tails}$</th>
<th>$a_{coop}$</th>
<th>$m_{heads}$</th>
<th>$m_{tails}$</th>
<th>$m_{coop}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{heads}$</td>
<td>0, 1</td>
<td>1, 0</td>
<td>0, 0</td>
<td>√</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>$t_{tails}$</td>
<td>1, 0</td>
<td>0, 1</td>
<td>0, 0</td>
<td>−</td>
<td>√</td>
<td>−</td>
</tr>
<tr>
<td>$t_{coop}$</td>
<td>0, 0</td>
<td>0, 0</td>
<td>1, 1</td>
<td>−</td>
<td>−</td>
<td>√</td>
</tr>
</tbody>
</table>

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 $m_{coop}$ if it is cooperation time and that she will use whatever other message, but certainly not $m_{coop}$, if it is not cooperation time; Bob, on the other hand, I would clearly expect to trust and believe message $m_{coop}$ and I would also expect him not to believe that either $m_{heads}$ or $m_{tails}$ 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
Chapter 1. What is Game Theoretic Pragmatics?

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. 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.