Rationalizable signaling

Gerhard Jäger
University of Bielefeld
Faculty of Linguistics and Literature
PF 10 01 31, 33615 Bielefeld, Germany
phone: +49-521-106-3576, fax: +49-521-106-2996
Gerhard.Jaeger@uni-bielefeld.de

April 2008

1 Introduction

Game theory is a branch of applied mathematics that deals with the analysis of strategic interaction. A game, in the sense of game theory, is characterized by the following properties: 1. There are at least two *players*. 2. The players interact, and the interaction results in a certain *outcome*. 3. Each player has a choice between various courses of action, their *strategies*. 4. The outcome of the interaction depends on the choice of strategy of each player. 5. Each player has a *preference ordering* over outcomes.

Preferences are usually encoded as numerical values, so-called *utilities* or *payoffs*, that are assigned to possible outcomes.

One of the objectives of game theory is to derive insights how rational players ought to behave in a strategic situation. A rational player is a player that holds some (possibly probabilistic) consistent beliefs about the structure of the game and the strategies of the other players, and that will choose their strategy in such a way that their expected utility is maximized. This entails, *inter alia*, that rational players do not have altruistic or spiteful motives. Also, rational players are assumed to be logically omniscient (they take all logical consequences of their beliefs into account in their decisions). In the standard interpretation of game theory, it is actually common knowledge among the players that all players are rational in this sense.

In many applications of game theory, communication between the players may affect the outcome of the game. Also, communication itself can be analyzed as a game. Therefore the game theoretic analysis of communication has attracted a good deal of attention in the literature.

As pointed out for instance by Stalnaker (2005), there is actually a strong affinity between rationalistic game theory and the kind of reasoning that is used in Gricean pragmatics (both in the sense of Grice's 1957 notion of non-natural meaning, and in the sense of Grice's 1975 concept of conversational implicatures). Stalnaker writes:

> "As many people have noticed, Gricean ideas naturally suggest a game theoretic treatment. The patterns of iterated knowledge and belief that are characteristic of game theoretic reasoning are prominent in Grice's discussions of speaker meaning, and the pattern of strategic reasoning that Grice discussed in the derivation of conversational implicatures are patterns that game theory is designed to clarify. ... [G]ame theory provides some sharp tools for formulating some of Grice's ideas[.] ... And I think Gricean ideas will throw some light on the problems game theorists face when they try to model communicative success."

This quotation succinctly summarizes the purpose of the present article. In the next section, I will briefly recapitulate some basic ideas about rational communication from the game theoretic literature. I will modify this model by incorporating some concepts from Gricean pragmatics informally in section 3, and I will present a formal framework for computing pragmatic interpretation from the literal meaning of expressions and the preferences and beliefs of the language users in section 4. Section 5 contains a series of examples that serve to illustrate the empirical predictions of this model and to compare it to other neo-Gricean theories like bidirectional Optimality Theory. Sections 6 and 7 contain pointers to related work and concluding remarks.

2 Signaling games

Let us consider a very elementary example for a situation where communication may make a difference. Suppose Sally pays Robin a visit, and Robin wants to offer his guest something to drink, either tea or coffee. Sally is either a tea drinker or a coffee drinker, and Robin prefers the outcome where Sally receives her favorite drink over the other outcome, but he does not know Sally's preferences.

We may formalize this scenario as follows: There are two possible worlds, w_1 and w_2. In w_1, Sally prefers tea; in w_2 she prefers coffee. Robin has a choice between two actions. Offering tea would be action a_1, and offering coffee is action a_2. Sally knows which world they are in, but Robin does not know it. Let us assume that Robin assigns both worlds an *a priori* probability of 50%. So the scenario can be represented by table 1. Rows represent possible worlds and columns represent Robin's actions. The first number in each cell gives Sally's payoff for this configuration, and the second number Robin's payoff.

Without any further coordination between the players, Robin will receive an expected payoff of 0.5 for either action, and hence Sally will also receive, on average, a payoff of 0.5. They can do better though if they communicate. Suppose it Robin expects that Sally says "tea" in w_1 and "coffee" in w_2. Then the rational course of action for Robin is to perform a_1 if he hears "tea", and to perform a_2 upon hearing "coffee". If Sally knows that Robin

	a_1	a_2
w_1	$1;1$	$0;0$
w_2	$0;0$	$1;1$

Table 1: A simple coordination scenario

will react to these signals in this way, it is in fact rational for her to say "tea" in w_1 and "coffee" in w_2.

So adding the option for communication may improve the payoff of both players. Technically, the original scenario (which is not really a game but a decision problem because Sally has no choice between actions) is transformed into a *signaling game.* Here the sender (Sally in the example) can send signals, and she can condition the choice of signals on the actual world. So a strategy for the sender is a function from possible worlds to signals. The receiver (Robin in the example) can condition his action on the signal received. So a strategy for the receiver is a function from signals to actions. (Analogous scenario's are studied extensively by Lewis 1969.)

The above example suggests that rational players will benefit from the option of communication. Things are not that simple though. If Sally says "I want tea" in w_1 and "I want coffee" in w_2, and Robin interprets "I want tea" as a_1 and "I want coffee" as a_2, both players benefit. Let us call this mapping from world to signals to actions L_1. They would receive the same benefit though if Sally said "I want coffee" in w_1 and "I want tea" in w_2, and Robin interprets "I want coffee" as a_1 and "I want tea" as a_2, which I will call L_2.[1] Pure reason does not provide a clue to decide between these two ways to coordinate. It is thus consistent with rationality that Sally assumes Robin to use L_2 and thus to signal according to L_2, while Robin assumes Sally to use L_1, and thus will interpret her signals according to L_1. In this situation, Robin will perform a_2 in w_1 and a_1 in w_2. Both players would receive the worst possible expected payoff of 0 here.

These considerations ignore the fact that the two signals do have a conventional meaning which is known to both players. L_1 is *a priori* much more plausible than L_2 because in L_1 Sally always says the truth, and Robin always believes the literal meaning of Sally's message.

Rational players cannot always rely on the honesty/credulity of the other player though. Consider the scenario from table 2 (unless otherwise indicated, I will always assume a uniform probability distribution over possible worlds):

Here the interests of Sally and Robin are strictly opposed; everybody can only win as much as the other one looses. Here too, there are two signals that Sally can send. We call them f_1 and f_2. They both have a conventional literal meaning: $[\![f_1]\!] = \{w_1\}$ and $[\![f_2]\!] = \{w_2\}$. If Robin is credulous, he will react to f_1 with a_2 and to f_2 with a_1. If Sally believes this and is rational, she will be dishonest and send f_1 in w_2 and f_2 in w_1. If Robin

[1] In the game-theoretic terminology, both L_1 and L_2 constitute strict Nash equilibria.

	a_1	a_2
w_1	$1; -1$	$-1; 1$
w_2	$-1; 1$	$1; -1$

Table 2: A simple zero sum scenario

is not quite so credulous, he may anticipate this and switch his strategy accordingly, etc. In fact, it turns out that with or without communication, any strategy is rationalizable in this game.[2] The lesson here is that communication might help in situations where the interests of the players are aligned, but it does not make a difference if these interests are opposed.

The most interesting scenarios, of course, are those where the interests of the players are partially, but not completely aligned. In a very influential paper, Crawford and Sobel (1982) showed that in such intermediate scenarios communication can be beneficial to both players. I will just discuss a simplified example here for illustration.

Suppose there are ten possible worlds, $w_1, \cdots, w_{10}$, and there are ten possible actions, $a_1, \cdots, a_{10}$. Sally prefers outcomes where the action index is one unit higher than the world index. More precisely, Sally's utility function u^S (as a function from worlds and actions to real numbers) can be given by

$$u^S(w_i, a_j) = -(i + 1 - j)^2$$

Robin, on the other hand, prefers scenarios where world index and action index are identical:

$$u^R(w_i, a_j) = -(i - j)^2$$

So Sally has an incentive to exaggerate the world index, but not too much. You can imagine that Sally is a job applicant, Robin is a potential employer, the world index represents Sally's level of skill, and the action index represents the kind of job that Robin is willing to give to Sally (with high indices representing highly demanding and well-paid jobs, and vice versa). Sally would prefer a situation where Robin slightly overestimates her skill level so that she gets a higher wage, but if she exaggerates too much, she might get a job that is much too demanding for her. Robin, on the other hand, prefers to hire Sally exactly according to her skill level.

It would seem that here, communication is not possible if both players are rational. Suppose Sally is in world w_2. Then she wants Robin to believe she is in w_3, and she will thus say f_3.[3] However, Robin will anticipate this and map f_3 to a_2. So Sally should actually say f_4 to induce a_3, but Robin might anticipate this as well, etc. After some more

[2] A strategy s is *rationalizable* if there is a consistent set of beliefs such that s maximizes the expected payoff of the player, given these beliefs and the assumption that rationality of all players is common knowledge.

[3] From now on, I will assume that $[\![f_i]\!] = \{w_i\}$ for all indices i unless otherwise stated.

rounds of reasoning, it will turn out that Sally will always say f_{10}, and Robin will ignore this and always choose a_5 or a_6 — the actions that maximizes his expected payoff in the absence of more specific information.

This reasoning crucially relies on the presence of signals $f_1, \cdots, f_{10}$ with a highly specific meaning. Suppose there are actually only two possible signals, f_{1-3} (with $[\![f_{1-3}]\!] = \{w_1, w_2, w_3\}$), and f_{4-10} (with $[\![f_{4-10}]\!] = \{w_4, \cdots, w_{10}\}$). Then it is actually rationalizable for Robin to believe the literal meaning of both messages, and to map f_{1-3} to a_2, and f_{4-10} to a_7. Also, it is rationalizable for Sally to use each of the two signals if and only if it is true. This pair of strategies — the honest sender strategy and the credulous receiver strategy — actually form a Nash equilibrium. This means that no player would have an incentive to change their strategy if they knew the other player's strategy.

This example illustrates two important insights: (a) it can be rational for both players to use communication even if their interests are not completely aligned,[4] and (b) whether or not it is rational to be honest/credulous may depend on the space of available messages. If Sally could use signals with a very specific meaning, this might tempt her into trying to deceive Robin, which, if anticipated, would lead to a breakdown of communication. If only sufficiently vague messages are available, this temptation does not arise here.

Table 3 represents another example that may illustrate this point. It is taken from Rabin (1990).

	a_1	a_2	a_3
w_1	$10; 10$	$0; 0$	$0; 0$
w_2	$0; 0$	$10; 10$	$5; 7$
w_3	$0; 0$	$10; 0$	$5; 7$

Table 3: Partially aligned interests

In w_1, Sally's and Robin's interests are identical; they both want Robin to take action a_1. So if Sally sends f_1, Robin has no reason to doubt it, and he will react to it by performing a_1. *Prima facie*, it might seem that the same holds for w_2. Here, both players would prefer Robin to take action a_2. However, in w_3 Sally also prefers Robin to take action a_2, while Robin would prefer a_3 if he knew that w_3 is the case. So in w_3 the interests of the players diverge, and Sally might be tempted to send the signal f_2 both in w_2 and w_3. Robin is thus well-advised not to believe f_2 in its literal meaning. If he does not know whether he is in w_2 or in w_3, his rational action is to hedge his bets and to perform a_3 after all, which guarantees him an expected payoff of 7 (against an expected payoff of 5 for a_2).

After performing these reasoning steps, Sally will perhaps convince herself that she has no chance to manipulate Robin into performing a_2. The best thing she can do both in w_2

[4] Without communication, the best thing Robin can do is choose either a_5 or a_6, which guarantees him a payoff of -8.5, while Sally would get a payoff of either -10.5 (for a_5) or -8.5 (for a_6). With communication in the described way, Robin's expected payoff rises to -4, and Sally's to -3.

and in w_3 is to prevent him from performing a_1. It is thus rational for Sally to say f_{23} (where $[\![f_{23}]\!] = \{w_2, w_3\}$) both in w_2 and w_3. So the situation that is most beneficial for both players is the one where only the signals f_1 and f_{23} are used, Sally uses each signal if and only if it is true, and Robin believes her and acts accordingly. In Rabin's terminology, f_1 is *credible* in this scenario, while f_2 would not be credible. Simplifying somewhat, a message f is credible (according to Rabin 1990) if it is rational for the sender to use it whenever it is true provided she can expect it to be believed, and if it is rational for the receiver to act as if she believed it provided the sender uses it whenever it is true.

Incidentally, f_{23} would not come out as credible in Rabin's model. The reason is that in w_2 or w_3, Sally might try to convince Robin that they are in w_2 and thus to induce action a_2. This is not possible via credible communication, but Sally might believe that she is capable to outsmart Robin by taking some other action.[5]

As was argued in the beginning, rationality alone is insufficient to coordinate players in such a way that signals receive a stable interpretation. This is even the case if signals do have a conventionalized meaning that is known to all players (as is the case for expressions from some natural language if both players know that language). Rabin proposes that, beyond being rational, reasonable sender will always send a true credible message if this is possible, and reasonable receivers will always believe any credible message.[6] In many cases, this reduces the space of rationalizable strategies significantly and thus ensures a certain amount of information transmission that is in the interest of both players.

3 Gricean reasoning

The kind of reasoning that was informally employed in the last section is reminiscent to pragmatic reasoning in the tradition of Grice (1975). First, information can only be exchanged between rational agents if it is in the good interest of both agents that this information transfer takes place. This intuition, which Grice captured in his Cooperative Principle, is implicit in the notion of credibility. Also, Rabin adopts a default assumption that messages are used according to their conventional meaning, unless overarching rationality considerations dictate otherwise. This corresponds to Grice's Maxim of Quality. Furthermore, the first part of the Maxim of Quantity — "make your contribution as informative as is required (for the current purpose of the exchange)" — is implicit in the notion of rationality. For instance, suppose Sally is in world w_1 in the scenario described in table 3. Then rationality requires her to transmit the information $\{w_1\}$ if there is a reliable way of doing so. If she would send a message which Robin would interpret as $\{w_1, w_2\}$, this would leave Robin in a state where he does not know whether a_1 or a_2 is the appropriate action. So sending an under-informative message would be irrational for Sally.

[5] Thanks to Michael Franke for pointing this out to me.

[6] In many scenarios, the intuitions about what constitutes a credible message is somewhat less clear than in the ones presented here. This has led to a lively debate about how credibility should be precisely defined. The interested reader is referred to Rabin (1992); Farrell (1993); Farrell and Rabin (1996); Zapater (1997); Stalnaker (2005) and the literature cited therein.

Despite these similarities, there are some crucial differences between Rabin's model and Gricean reasoning. To illustrate this, let us consider a schematic example of a scalar implicature. The utility structure is given in table 4. Suppose, as before, that there are three messages, f_1, f_2 and f_{12}, with the conventionalized meanings $\{w_1\}$, $\{w_2\}$, and $\{w_1, w_2\}$ respectively. However, we now assume that sending a message may incur some costs for the sender, and that different messages incur different costs. In the specific example, we assume that $c(f_1) = c(f_{12}) = 0$, and $c(f_2) = 2$, where $c(f)$ is the cost that the sender has to pay for sending message f. So the sender's utility is now a three-place function u^S that depends on the actual world, the message sent, and the action that the receiver takes. If $v^S(w, a)$ is the distribution of sender payoffs that is given in table 4 above, the sender's overall utility is

	a_1	a_2	a_3
w_1	$10, 10$	$0, 0$	$9, 9$
w_2	$0, 0$	$10, 10$	$9, 9$

Table 4: context c_1: scalar implicature

$$u^S(w, f, a) = v^S(w, a) - c(f)$$

You can imagine that Robin wants to know who was at the party last night, and Sally knows the answer. In w_1, all girls were at the party, and in w_2 some but not all girls were there. f_1 is the message "All girls were at the party", f_2 is "Some but not all girls were at the party", and f_{12} is "Some girls were at the party." Obviously f_2 is more complex than the other two messages, which are approximately equally complex. This is covered by the assignment of costs.

According to Gricean pragmatics, Sally would reason roughly as follows: If I am in w_1, I want Robin to perform a_1 because this gives me a utility of 10. a_1 is what he would do if he believed that he is in w_1. I can try to convince him of this fact by saying f_1. It is not advisable to say f_2, because if Robin believed it, he would perform a_2, which gives me a utility of a mere -2. Also saying f_{12} is not optimal because if Robin believes it, he will perform a_3, leading to a utility of 9. So it seems reasonable to send f_1 in w_1.

If we are in w_2, it might seem reasonable to say f_2 because if Robin believes it, he will perform a_2, which is my favorite outcome. However, I will have to pay the costs of 2, so my net utility is only 8. If I say f_{12} and Robin believes it, he will perform a_3. As f_{12} is costless for me, my net utility is 9, which is better than 8. So in w_2 I will send f_{12}.

Robin in turn will anticipate that Sally will reason this way. If he is confronted with the message f_1, he will infer that he is in w_1, and he will perform a_1. If he hears f_{12}, he will infer that w_2 is the case, and he will perform a_2 after all.

Sally, being aware of this fact, will reason: This taken into consideration, it is even more beneficial for me to send f_{12} if I am in w_2 because this will give me the maximal payoff of 10. So I have no reason to change the plan of sending f_1 in w_1 and f_{12} in w_2.

This reasoning leads to a sender strategy where f_{12} is sent if and only if $\{w_2\}$ is true. Following Lewis (1969), we will call the set of worlds where a certain message is sent its *indicative meaning* (as opposed to its imperative meaning, which is the set of actions that the receiver might perform upon receiving that message). In our example, the indicative meaning of f_{12} thus turns out to be $\{w_2\}$, which is a proper subset of its literal meaning $\{w_1, w_2\}$. The information that w_1 is not the case is a scalar implicature — "some" is pragmatically interpreted as "some but not all."

As in the examples discussed in the previous section, the inferences that are used here start with a default assumption that messages are used according to their literal interpretation, but this is only a provisional assumption that is adopted if this is not in contradiction with rationality. Nevertheless, there are crucial differences. In the ultimate outcome that is inferred, f_{12} would not count as credible in the sense of Rabin, because in w_1 it is literally true, but Sally would nevertheless not send it. Likewise, f_2 is not credible in the technical sense because it is not rational for Sally to send it, even if it is true and if Robin would believe it.

The reasoning pattern that is used here makes implicit use of the notion of the *best response* of a player to a certain probabilistic belief. A best response (that need not be unique) to a belief state is a strategy that maximizes the expected payoff of the player as compared to all strategies at their disposal, given this belief state. Rational players will always play some best response to their beliefs.

Suppose an external observer (that might be Robin, Sally who tries to figure out Robin's expectations, Robin who tries to figure out Sally's expectations about his intentions etc., or we as modelers) has some partial knowledge about Sally's belief state. There is some set of receiver strategies R, and the observer knows that Sally expects Robin to play some strategy of R, and that Sally cannot exclude any element of R for sure. The observer does not know which probability Sally assigns to the elements of R. Then any probability distribution of R (that only assigns positive probabilities to elements of R, to be precise) is a possible belief state of Sally's, as far as the observer's knowledge is concerned. Hence any best response of Sally's to such a belief state is a *potential best response* for Sally against R. All the observer can predict with certainty if he assumes Sally to be rational is that she will play some potential best response against R. (Since Sally holds a specific private belief, she will actually only consider a subset of the potential best responses, but the observer does not know which one.)

The iterative inference process that was used in the computation of the implicature above can be informally described as follows:

- Sally provisionally assumes that Robin is entirely credulous, and that he conditions his actions only on the literal interpretation of the message received. Let us call the set[7] of credulous strategies R_0. In the first round of reasoning, Sally might ponder any strategy that is a potential best response against R_0. Let us call this set of strategies S_0.

[7] There might be more than one credulous strategy because several actions may yield the same maximal payoff for Robin in certain situations.

- In the next round, Robin might ponder all strategies that are potential best responses against S_0. The set of these strategies is R_1.
- ...
- S_n (R_{n+1}) is the set of strategies that are potential best responses against R_n (S_n).
- If a certain strategy S (R) cannot be excluded by this kind of reasoning (i.e. if there are infinitely many indices i such that $S \in S_i$ ($R \in R_i$)), then S (R) is a *pragmatically rationalizable strategy.*

In the example, the scalar implicature arises because the difference between $v^S(w_2, a_2)$ and $v^S(w_2, a_3)$ is smaller than the costs of sending f_2. Suppose the utilities would be as in table 5, rather than as in table 4. Then the pragmatically rationalizable outcome would be that Sally uses f_2 in w_2, while f_{12} would never be used. Informally speaking, the reasoning here relies on a tension between the Maxim of Quantity and the Maxim of Manner. The implicature only arises if the utilities are such that Manner wins over Quantity.

	a_1	a_2	a_3
w_1	$10, 10$	$0, 0$	$6, 6$
w_2	$0, 0$	$10, 10$	$6, 6$

Table 5: context c_2: no scalar implicature

In a more realistic scenario, Robin might actually not know for sure what Sally's precise preferences are. If we call the utility matrix in table 4 *context* c_1[8], and the utilities in table 5 context c_2, Robin might hold some probabilistic belief about whether Sally is in c_1 or in c_2. Likewise, Sally need not know for sure which context Robin is in. Now in each round of the iterative reasoning process, the players will ponder each strategy that is a potential best response to any probability distribution over contexts and strategies in the previous round.[9]

Sally's reasoning will now start as follows: In w_1, I will definitely send f_1, no matter which context I am in. If I am in context c_1, it is better to send f_{12} if I am in w_2 because the costs of sending the more explicit message f_2 exceed the potential benefits. If I am in c_2 and w_2, however, it is advisable to use f_2.

Robin, in turn, will reason: If I hear f_1, we are definitely in w_1, and the best thing I can do is to perform a_1, no matter which context we are in. If I hear f_2, we are in c_2/w_2, and I will perform a_2. If I hear f_{12}, we are in c_1/w_2, and I will also play a_2.

[8] I use the term "context" in such a way here that the preferences of the players may vary between contexts (as well as between worlds), while the literal meaning of messages is invariant between contexts. So this notion of context has nothing to do with the knowledge state of the discourse participants or the interpretation of indexical expressions.

[9] Epistemically speaking, this means that I do not assume any common belief about which context the players are in, even though they might hold private beliefs.

So in S_1 Sally will infer: f_1 will induce a_1, and both f_2 and f_{12} will induce a_2, no matter which context Robin is in. Since f_{12} is less costly than f_2, I will always use f_1 in w_1 and f_{12} in w_2, regardless of the context I am in. Robin, in R_1, will thus conclude that his best response to f_1 is always a_1, and his best response to f_{12} is a_2. Nothing will change in later iterations. So here, the scalar implicature from "some" to "some but not all" will arise in all contexts, even though context c_2 by itself would not license it.

One might argue that this is not quite what happens in natural language use. Here we predict that f_2 would never be used. A more realistic outcome would be that f_2 is still interpreted as $\{w_2\}$, and that by using it, Sally conveys the message that it is very important to her that w_1 is in fact excluded.

What I believe is going on here is that there are also contexts where Sally does not know for sure which world she is in. In this case f_{12} might be sent in w_1 after all. Whether or not Robin derives the implicature in question would depend then on how much probability he assigns to this option.

To keep things simple, I will confine the technical model to be derived in this article to scenarios where the sender has complete factual knowledge, i.e. where she knows the identity of the actual world. A generalization to games where both players have incomplete information is certainly possible though.

4 The formal model

In this section I will develop a formal model that captures the intuitive reasoning from the last section.

A *semantic game* is a game between two players, the sender S and the receiver R. It is characterized by a set of contexts $\mathcal{C}$, a set of worlds $\mathcal{W}$, a set of signals $\mathcal{F}$, a set of actions $\mathcal{A}$, a probability distribution p^*, an interpretation function $[\![\cdot]\!]$, and a pair of utility functions u^S and u^R. In the context of this paper, I will confine the discussion to games where $\mathcal{C}$, $\mathcal{W}$, $\mathcal{F}$, and $\mathcal{A}$ are all finite. p^* is a probability distribution over $\mathcal{W}$. Intuitively, $p^*(w)$ is the *a priori* probability that the actual world is w. We assume that $p^*(w) > 0$ for all $w \in \mathcal{W}$.

$u^S \in \mathcal{C} \times \mathcal{W} \times \mathcal{F} \times \mathcal{A} \mapsto \mathbb{R}$ is the sender's utility function. There is some function $v^S \in \mathcal{C} \times \mathcal{W} \times \mathcal{A} \mapsto \mathbb{R}$ and some function $c \in \mathcal{F} \mapsto \mathbb{R}$ such that

$$u^S(c, w, f, a) = v^S(c, w, a) - c(f).$$

$u^R \in \mathcal{C} \times W \times \mathcal{A} \mapsto \mathbb{R}$ is the receiver's utility function. $[\![\cdot]\!] \in \mathcal{F} \mapsto \wp(W)$ is the semantic interpretation function that maps signals to propositions.

The space of pure sender strategies $\mathcal{S} = \mathcal{C} \times \mathcal{W} \mapsto \mathcal{F}$ is the set of functions from context/world pairs to signals. The space of pure receiver strategies $\mathcal{R} = \mathcal{C} \times \mathcal{F} \mapsto \mathcal{A}$ is the set of functions from context/signals pairs to actions.

The structure of the game is common knowledge between the players.

Some auxiliary notations: If M is a finite and non-empty set, $\Delta(M)$ is defined as

$$\Delta(M) = \{q \in M \mapsto [0,1] | \sum_{x \in M} q(x) = 1\}.$$

This is the set of probability distributions over M. A related notion is:

$$\mathrm{int}(\Delta(M)) = \{q \in M \mapsto (0,1] | \sum_{x \in M} q(x) = 1\}.$$

This is the set of probability distributions over M where each element of M receives a positive probability. The difference is subtle but important. Both $\Delta(\cdot)$ and $\mathrm{int}(\Delta(\cdot))$ can be used to model probabilistic beliefs. If we say that a player holds a belief from $\Delta(\mathcal{C})$, say, this means that they may exclude some contexts with absolute certainty. On the other hand, if Sally believes that Robin plays his strategy according to $\mathrm{int}(\Delta(R))$ for some set $R \subseteq \mathcal{R}$, then Sally may have certain guesses, but she is not able to exclude any strategy from R with certainty. We will use this to capture the intuition that the players may have biases, but they do not have other sources of established beliefs about the intentions of the other players beyond the assumption that pragmatic rationality is common knowledge.

Definition 1 *Let $\phi \subseteq \mathcal{W}$ be a proposition and $p \in \mathrm{int}(\Delta(\mathcal{W}))$ be a probability distribution over worlds.*

$$\begin{array}{rcl} A^*(c, \phi, p) & \doteq & \{a^* \in \mathcal{A} | a^* \in \arg_{a \in \mathcal{A}} \max \sum_{w \in \phi} p(w) u^R(c, w, a)\} \\ A^*(c, \phi) & \doteq & A^*(c, \phi, p^*) \end{array}$$

So $A^*(c, \phi, p)$ is the set of actions that might be optimal for the receiver if he is in context c, his (probabilistic) prior belief about the possible worlds is p, and this prior belief is updated with the information that he is in ϕ. $A^*(c, \phi)$ is the set of actions that the receiver believes to be optimal in c if he updates the prior belief p^* with ϕ.

The central step in the iterative process described above is the computation of the set of strategies that maximize the expected payoff of a player against some probability distribution over contexts and strategies of the other player. The notion of a *best response* captures this.

Definition 2

- *Let $r^* \in \mathcal{R}$ be a receiver strategy, $\sigma \in \Delta(\mathcal{S})$ a probability distribution over $\mathcal{S}$, and $q \in \Delta(\mathcal{C})$ a probability over contexts. (σ, q) represent a belief of the receiver.*

  $$r^* \in BR(\sigma, q)$$

  (r^* is a best response *of the receiver to (σ, q)) iff*

  $$\forall c \in \mathcal{C} : r^* \in \arg_{r \in \mathcal{R}} \max \sum_{s \in \mathcal{S}} \sigma(s) \sum_{c' \in \mathcal{C}} q(c') \sum_{w \in \mathcal{W}} p^*(w) u^R(c, w, r(c, s(c', w)))$$

- *Let $s^* \in \mathcal{S}$ a sender strategy, $\rho \in \Delta(\mathcal{R})$ a probability distribution over $\mathcal{R}$, and $q \in \Delta(\mathcal{C})$ a probability over contexts. (ρ, q) represent a belief of the sender.*

  $$s^* \in BR(\rho, q)$$

  (s^* is a best response *of the sender to (ρ, q)) iff*

  $$\forall c \in \mathcal{C} \forall w \in \mathcal{W} : s^* \in \arg_{s \in \mathcal{S}} \max \sum_{r \in \mathcal{R}} \rho(r) \sum_{c' \in \mathcal{C}} q(c') u^S(c, w, s(c, w), r(c', s(c, w)))$$

The set of *potential best responses* against some set P of strategies of the opposing player is the set of strategies that are best responses to some belief state that assigns positive probability exactly to the elements of P.

Definition 3

- *Let $S \subseteq \mathcal{S}$ be a set of sender strategies. The set of* potential best responses to S *is defined as*

  $$PBR(S) = \bigcup_{\sigma \in \text{int}(\Delta(S))} \bigcup_{q \in \Delta(\mathcal{C})} r \in BR(\sigma, q)$$

- *Let $R \subseteq \mathcal{R}$ be a set of receiver strategies. The set of* potential best responses to R *is defined as*

  $$PBR(R) = \bigcup_{\rho \in \text{int}(\Delta(R))} \bigcup_{q \in \Delta(\mathcal{C})} BR(\rho, q)$$

Suppose we know that Sally knows which context and world she is in, she believes for sure that Robin will play a strategy from R, and there is no more specific information that she believes to know for sure. We do not know which strategy from R Sally expects Robin to play with which likelihood, and which context Sally believes to be in. Under these conditions, all we can predict for sure is that Sally will play some strategy from $PBR(R)$ if she is rational.

The same seems to hold if we only know that Robin expects Sally to play some strategy from S. Then we can infer that Robin, if he is rational, will certainly play a strategy from $PBR(S)$. However, we may restrict his space of reasonable strategies even further. Suppose none of the strategies in S ever make use of the signal f. (Formally put, $f \in \mathcal{F} - \bigcup_{s \in S} \mathit{range}(s)$.) Then it does not make a difference how Robin would react to f, but he has to decide about the imperative meaning of f nevertheless (because receiver strategies are **total** functions from context/form pairs to actions). It seems reasonable to demand (and it leads to reasonable predictions, as we will see below) that Robin should, in the absence of evidence to the contrary, still assume that f is true. For instance, if Sally speaks English to Robin, and she suddenly throws in a sentence in Latin (that Robin happens to understand), Robin will probably assume that the Latin sentence is true, even if he did not expect her to use Latin.

If Robin encounters such an unexpected signal, he will have to revise his beliefs. In the previous paragraph I argued that this belief revision should result in an epistemic state where f is true. However, no further restrictions on Robin's belief revision policy will be stated. In particular, we will not demand that Robin will fall back to $p(\cdot|[\![f]\!])$, i.e. to the result of updating his prior belief with the literal interpretation of f. Robin will have to figure out an explanation why Sally used f despite his expectations to the contrary, and this explanation can bias his prior beliefs in any conceivable way. We have to assume though that the result of this believe revision is a consistent belief state, and that Robin will act rationally according to his new beliefs.

We can now proceed to define the iterative reasoning procedure that was informally described in the previous section.

Definition 4

$$\begin{aligned}
R_0 &\doteq \{r \in \mathcal{R} | \forall c \in \mathcal{C} \forall f \in \mathcal{F} : r(c,f) \in A^*(c, [\![f]\!])\} \\
S_n &\doteq PBR(R_n) \\
R_{n+1} &\doteq \{r \in PBR(S_n)| \\
&\quad\ \forall f \in \mathcal{F} - \bigcup_{s \in S_n} range(s) \forall c \in \mathcal{C} \exists p \in \mathrm{int}(\Delta(\mathcal{W})) : r(c,f) \in A^*(c, [\![f]\!], p)\}
\end{aligned}$$

R_0 is the set of credulous strategies of the receiver. S_n is the set of potential best responses of the sender against R_n. Likewise, R_{n+1} is the set of potential best responses of the receiver if he assumes that the sender plays a strategy from S_n in which he always tries to make sense of unexpected messages under the assumption that they are literally true.

The sets of *pragmatically rationalizable strategies* (PRS) are the set of sender strategies and receiver strategies that cannot be excluded for sure by the iterative reasoning process, no matter how deeply the reasoning goes.

Definition 5 $(\mathbf{S}, \mathbf{R}) \in \wp(\mathcal{S}) \times \wp(\mathcal{R})$*, the sets of* pragmatically rationalizable strategies*, are defined as follows:*

$$\begin{aligned}
\mathbf{S} &\doteq \{s \in \mathcal{S} | \forall n \in \mathbb{N} \exists m > n : s \in S_m\} \\
\mathbf{R} &\doteq \{r \in \mathcal{R} | \forall n \in \mathbb{N} \exists m > n : s \in R_m\}
\end{aligned}$$

Note that there are only finitely many strategies in $\mathcal{S}$ and $\mathcal{R}$ (because we are only considering pure strategies). Therefore there are only finitely many subsets thereof. The step from (S_n, R_n) to (S_{n+1}, R_{n+1}) is always deterministic. It follows that the iterative procedure will enter a cycle at some point, i.e. there are n^* and i^* such that for all $m > n^*$ and for all k: $(S_m, R_m) = (S_{m+k \cdot i^*}, R_{m+k \cdot i^*})$. This ensures that $(\mathbf{S}, \mathbf{R})$ is always defined.

As was mentioned above, a strategy is called rationalizable iff a rational player might use it, provided it is common knowledge that all players are rational. The following formal definition (adapted from Osborne 2003:383) is provably equivalent to this informal characterization:

Definition 6 *The strategy pair* $(s^*, r^*) \in \mathcal{S} \times \mathcal{R}$ *is* rationalizable *iff there exist sets* $S \subseteq \mathcal{S}$ *and* $R \subseteq \mathcal{R}$ *such that*

- $S \subseteq \bigcup_{\rho\in\Delta(R)} \bigcup_{q\in\Delta(\mathcal{C})} BR(\rho, q)$
- $R \subseteq \bigcup_{\sigma\in\Delta(S)} \bigcup_{q\in\Delta(\mathcal{C})} BR(\sigma, q)$
- $(s^*, r^*) \in S \times R$

In words, s^* and r^* are rationalizable iff they are elements of some sets S and R such that every element of S is a best response to some belief of the sender that only considers strategies in R possible, and every element of R is a best response to some belief of the receiver that only considers strategies in S possible.

The set of pragmatically rationalizable strategies are in fact rationalizable:

Theorem 1 *For all* $(s^*, r^*) \in \mathbf{S} \times \mathbf{R}$*:* (s^*, r^*) *is rationalizable.*

Proof: As there are only finitely many subsets of $\mathcal{S}$ and $\mathcal{R}$ and (S_{n+1}, R_{n+1}) is a function of (S_n, R_n) for all n, there must be some $m^* \geq 0, i^* > 0$ such that for all $k, l \geq 0 : (S_{m^*+k\cdot i^*+l}, R_{m^*+k\cdot i^*+l}) = (S_{m^*}, R_{m^*})$. Let $s \in \mathbf{S}$. Then there must be some l^* such that $s \in S_{m^*+l^*} = PBR(R_{m^*+l^*})$ and $R_{m^*+l^*} \subseteq \mathbf{R}$. So there are $\rho \in \mathrm{int}(\Delta(R_{m^*+l^*}))$ and $q \in \Delta(\mathcal{C})$ such that $s \in BR(\rho, q)$. Trivially, ρ can be extended to some $\rho' \in \Delta(\mathbf{R})$ by assigning zero probability to all elements of $\mathbf{R} - R_{m^*+l^*}$ such that $BR(\rho, q) = BR(\rho', q)$. So $s \in \bigcup_{\rho\in\Delta(\mathbf{R})} \bigcup_{q\in\Delta(\mathcal{C})} BR(\rho, q)$.

In a similar way, suppose $r \in \mathbf{R}$. Then there must be some l^* sucht that $r \in R_{m^*+l^*} = R_{m^*+i^*+l^*} \subseteq PBR(S_{m^*+i^*+l^*-1})$ and $S_{m^*+i^*+l^*-1} \subseteq \mathbf{S}$. By an argument analogous to the previous case, it follows that $r \in \bigcup_{\sigma\in\Delta(\mathbf{S})} \bigcup_{q\in\Delta(\mathcal{C})} BR(\sigma, q)$. Hence

$$\mathbf{S} \subseteq \bigcup_{\rho\in\Delta(\mathbf{R})} \bigcup_{q\in\Delta(\mathcal{C})} BR(\rho, q)$$

and

$$\mathbf{R} \subseteq \bigcup_{\sigma\in\Delta(\mathbf{S})} \bigcup_{q\in\Delta(\mathcal{C})} BR(\sigma, q).$$

So any element of $\mathbf{S} \times \mathbf{R}$ is rationalizable. ⊣

5 Examples

In the light of this formal definition, let us consider some of the previous examples again, which are repeated here for convenience.

	a_1	a_2
w_1	$1;1$	$0;0$
w_2	$0;0$	$1;1$

Table 6: Example 1

Example 1 Completely aligned interests: We assume that for all signals $f : c(f) = 0$. There is only one context; v^S and u^R are given in table 6.

Here is the sequence of iterated computation of potential best responses, starting with the set R_0 of credulous strategies. The representation should be self-explanatory; every function that pairs one of the arguments in the left column with one of the arguments in the right column is part of the strategy set in question.

$$\mathbf{R} = R_0 = \left[\begin{array}{lll} f_1 & \to & a_1 \\ f_2 & \to & a_2 \\ f_{12} & \to & a_1/a_2 \end{array}\right]$$

$$\mathbf{S} = S_0 = \left[\begin{array}{lll} w_1 & \to & f_1 \\ w_2 & \to & f_2 \end{array}\right]$$

Example 2 Completely opposing interests: We still assume "cheap talk", i.e. all messages are costless. The utilities are repeated in table 7

	a_1	a_2
w_1	$1;-1$	$-1;1$
w_2	$-1;1$	$1;-1$

Table 7: Example 2

Here the iterative procedure enters a never-ending cycle:

$$
\begin{aligned}
R_0 &= \begin{bmatrix} f_1 & \to & a_2 \\ f_2 & \to & a_1 \\ f_{12} & \to & a_1/a_2 \end{bmatrix} \\
S_0 &= \begin{bmatrix} w_1 & \to & f_2 \\ w_2 & \to & f_1 \end{bmatrix} \\
R_1 &= \begin{bmatrix} f_1 & \to & a_1 \\ f_2 & \to & a_2 \\ f_{12} & \to & a_1/a_2 \end{bmatrix} \\
S_1 &= \begin{bmatrix} w_1 & \to & f_1 \\ w_2 & \to & f_2 \end{bmatrix} \\
R_2 &= R_0 \\
S_2 &= S_0 \\
&\vdots \\
\mathbf{R} &= \begin{bmatrix} f_1/f_2 & \to & a_1/a_2 \end{bmatrix} \\
\mathbf{S} &= \begin{bmatrix} w_1/w_2 & \to & f_1/f_2/f_{12} \end{bmatrix}
\end{aligned}
$$

So if the interests of the players are completely opposed, no communication will ensue.

Example 3 Rabin's example with partially aligned interests; the utilities are as in table 8 and all signals are costless.

	a_1	a_2	a_3
w_1	$10; 10$	$0; 0$	$0; 0$
w_2	$0; 0$	$10; 10$	$5; 7$
w_3	$0; 0$	$10; 0$	$5; 7$

Table 8: Example 3

$$
\begin{array}{rcl}
R_0 & = & \left[\begin{array}{lcl} f_1/f_{13} & \rightarrow & a_1 \\ f_2 & \rightarrow & a_2 \\ f_3/f_{23}/f_{123} & \rightarrow & a_3 \\ f_{12} & \rightarrow & a_1/a_2 \end{array}\right] \\
S_0 & = & \left[\begin{array}{lcl} w_1 & \rightarrow & f_1/f_{13} \\ w_2/w_3 & \rightarrow & f_2 \end{array}\right]
\end{array}
$$

$$
\begin{array}{rcl}
R_1 & = & \left[\begin{array}{lcl} f_1/f_{13} & \rightarrow & a_1 \\ f_2/f_3 & \rightarrow & a_3 \\ f_{12} & \rightarrow & a_1/a_2 \\ f_{23} & \rightarrow & a_2/a_3 \\ f_{123} & \rightarrow & a_1/a_2/a_3 \end{array}\right] \\
S_1 & = & \left[\begin{array}{lcl} w_1 & \rightarrow & f_1/f_{13} \\ w_2/w_3 & \rightarrow & f_{12}/f_{23}/f_{123} \end{array}\right]
\end{array}
$$

$$
\begin{array}{rcl}
R_2 & = & \left[\begin{array}{lcl} f_1/f_{13} & \rightarrow & a_1 \\ f_2 & \rightarrow & a_2 \\ f_3 & \rightarrow & a_3 \\ f_{12}/f_{23}/f_{123} & \rightarrow & a_2/a_3 \end{array}\right] \\
\neg(R_2(f_{12}) = & R_2(f_{23}) & = R_2(f_{123}) = a_2)
\end{array}
$$

$$
\begin{array}{rcl}
S_2 & = & S_0 \\
R_3 & = & R_1 \\
 & \vdots & \\
\mathbf{R} & = & R_1 \cup R_2 \\
\mathbf{S} & = & S_0 \cup S_1
\end{array}
$$

Note that no stable communication will emerge here in w_2 and w_3. Starting in S_0, Sally has the same set of options in w_2 and w_3. She may or may not choose to differentiate between w_2 and w_3; there are some potential best responses against R_1 that do and some that do not. Depending on Robin's private belief, he may expect to be able to differentiate between w_2 and w_3 on the basis of Sally's signal (and thus react to some signals with a_2), or he may prefer to play safe and choose a_3.

The situation changes drastically if the set of signals is confined to f_1 and f_{23}. Then we have

$$
\begin{aligned}
\mathbf{R} = R_0 &= \left[\begin{array}{lcl} f_1 & \to & a_1 \\ f_{23} & \to & a_3 \end{array}\right] \\
\mathbf{S} = S_0 &= \left[\begin{array}{lcl} w_1 & \to & f_1 \\ w_2/w_3 & \to & f_{23} \end{array}\right].
\end{aligned}
$$

Example 4 Next we will reconsider the example of the scalar implicature discussed above. Now we have two contexts, c_1 and c_2. The utilities are given in table 9.

c_1 :

	a_1	a_2	a_3
w_1	$10; 10$	$0; 0$	$6; 6$
w_2	$0; 0$	$10; 10$	$6; 6$

c_2 :

	a_1	a_2	a_3
w_1	$10; 10$	$0; 0$	$9; 9$
w_2	$0; 0$	$10; 10$	$9; 9$

Table 9: Example 4

The signaling costs are as follows: $c(f_1) = c(f_{12}) = 0$ and $c(f_2) = 2$.

$$
\begin{aligned}
R_0 &= \left[\begin{array}{lcl} (c_1, f_1)/(c_2, f_1) & \to & a_1 \\ (c_1, f_2)/(c_2, f_2) & \to & a_2 \\ (c_1, f_{12})/(c_2, f_{12}) & \to & a_3 \end{array}\right] \\
S_0 &= \left[\begin{array}{lcl} (c_1, w_1)/(c_2, w_1) & \to & f_1 \\ (c_1, w_2) & \to & f_2 \\ (c_2, w_2) & \to & f_{12} \end{array}\right] \\
\mathbf{R} = R_1 &= \left[\begin{array}{lcl} (c_1, f_1)/(c_2, f_1) & \to & a_1 \\ (c_1, f_2)/(c_2, f_2)/(c_1, f_{12})/(c_2, f_{12}) & \to & a_2 \end{array}\right] \\
\mathbf{S} = S_1 &= \left[\begin{array}{lcl} (c_1, w_1)/(c_2, w_1) & \to & f_1 \\ (c_1, w_2)/(c_2, w_2) & \to & f_{12} \end{array}\right]
\end{aligned}
$$

The previous example illustrated how pragmatic rationalizability formalizes the intuition behind Levinson's (2000) **Q-Heuristics** "What isn't said, isn't." This heuristics accounts, *inter alia* for scalar and clausal implicatures like the following:

(1) a. Some boys came in. $\rightsquigarrow$ Not all boys came in.

b. Three boys came in. $\rightsquigarrow$ Exactly three boys came in.

(2) a. If John comes, I will leave. $\rightsquigarrow$ It is open whether John comes.

b. John tried to reach the summit. $\rightsquigarrow$ John did not reach the summit.

The essential pattern here is as in the schematic example above: There are two expressions A and B of comparable complexity such that the literal meaning of A entails the literal meaning of B. There is no simple expression for the concept "B but not A". In this scenario, a usage of "B" will implicate that A is false.

Example 5 Levinson assumes two further pragmatic principles that, together with the Q-principle, are supposed to replace Grice's maxims in the derivation of generalized conversational implicatures. The second heuristics, called **I-Heuristics**, says: "What is simply described is stereotypically exemplified." It accounts for phenomena of pragmatic strengthening, as illustrated in the following examples:

(3) a. John's book is good. $\rightsquigarrow$ The book that John is reading or that he has written is good.

b. a secretary $\rightsquigarrow$ a female secretary

c. road $\rightsquigarrow$ hard-surfaced road

The notion of "stereotypically exemplification" is somewhat vague and difficult to translate into the language of game theory. I will assume that propositions with a high prior probability are stereotypical. Also, I take it that "simple description" can be translated into "low signaling costs." So the principle amounts to "Likely propositions are expressed by cheap forms."

Let us construct a schematic example of such a scenario. Suppose there are two possible worlds (which may also stand for objects, like a hard surfaced vs. soft-surfaced road) w_1 and w_2, such that w_1 is *a priori* much more likely than w_2. Let us say that $p(w_1)/p(w_2) = 3$. There are three possible actions for Robin; he may choose a_1 if he expects w_1 to be correct, a_2 if he expects w_2, and a_3 if he finds it too risky to choose.

There are again three signals, f_1, f_2 and f_{12}. This time the more general expression f_{12} (corresponding for instance to "road") is cheap, while the two specific expressions f_1 and f_2 ("hard-surfaced road" and "soft-surfaced road") are more expensive: $c(f_1) = c(f_2) = 5$, and $c(f_{12}) = 0$.

The interests of Sally and Robin are completely aligned, except for the signaling costs which only matter for Sally. There are three contexts. In c_1 and c_2, it is safest for Robin to choose a_3 if he decides on the basis of the prior probability. In c_3 it makes sense to choose either a_1 if he only knows the prior probabilities because the payoff of a_3 is rather low (but still higher than making the wrong choice between a_1 and a_2). In c_1, but not in c_2 it would be rational for Sally to use a costly message if this is the only way to make Robin perform a_1 rather than a_3. The precise utilities are given in table 10.

		a_1	a_2	a_3
c_1 :	w_1	$28;28$	$0;0$	$22;22$
	w_2	$0;0$	$28;28$	$22;22$

		a_1	a_2	a_3
c_2 :	w_1	$28;28$	$0;0$	$25;25$
	w_2	$0;0$	$28;28$	$25;25$

		a_1	a_2	a_3
c_3 :	w_1	$28;28$	$0;0$	$10;10$
	w_2	$0;0$	$28;28$	$10;10$

Table 10: Example 5

$$R_0 \quad = \quad \left[\begin{array}{lcl} (c_1, f_1)/(c_2, f_1)/(c_3, f_1)/(c_3, f_{12}) & \rightarrow & a_1 \\ (c_1, f_2)/(c_2, f_2)/(c_3, f_2) & \rightarrow & a_2 \\ (c_1, f_{12})/(c_2, f_{12}) & \rightarrow & a_3 \end{array} \right]$$

$$\mathbf{S} = S_0 \quad = \quad \left[\begin{array}{lcl} (c_1, w_1)/(c_3, w_1) & \rightarrow & f_1/f_{12} \\ (c_1, w_2)/(c_3, w_2) & \rightarrow & f_2 \\ (c_2, w_1) & \rightarrow & f_{12} \\ (c_2, w_2) & \rightarrow & f_2/f_{12} \end{array} \right]$$

$$\mathbf{R} = R_1 \quad = \quad \left[\begin{array}{lcl} (c_1, f_1)/(c_2, f_1)/(c_3, f_1)/(c_3, f_{12}) & \rightarrow & a_1 \\ (c_1, f_2)/(c_2, f_2)/(c_3, f_2) & \rightarrow & a_2 \\ (c_1, f_{12})/(c_2, f_{12}) & \rightarrow & a_1/a_3 \end{array} \right]$$

Here both f_1 and f_2 retain its literal meaning under pragmatic rationalizability. The unspecific f_{12} also retains its literal meaning in c_2. In c_1 and c_3, though, its meaning is pragmatically strengthened to $\{w_1\}$. Another way of putting is to say that f_{12} is *pragmatically ambiguous* here. Even though it has an unambiguous semantic meaning, its pragmatic interpretation varies between contexts. It is noteworthy here that f_{12} can never be strengthened to mean $\{w_2\}$. Applying it to the example, this means that a simple non-specific expression like “road” can either retain its unspecific meaning, or it can be pragmatically strengthened to its stereotypical instantiation (like *hard-surfaced road* here). It can never be strengthened to a non-stereotypical meaning though.

Example 6 Levinson's third heuristics is the **M-heuristics**: "What is said in an abnormal way isn't normal." It is also known, after Horn (1984), as **division of pragmatic labor**. A typical example is the following:

(4) a. John stopped the car.

b. John made the car stop.

The two sentences are arguably semantically synonymous. Nevertheless they carry different pragmatic meanings if uttered in a neutral context. (4a) is preferably interpreted as *John stopped the car in a regular way, like using the foot brake.* This would be another example for the I-heuristics. (4b), however, is also pragmatically strengthened. It means something like *John stopped the car in an abnormal way, like driving it against a wall, making a sharp u-turn, driving up a steep mountain, etc.*

This can be modeled quite straightforwardly. Suppose there are again two worlds, w_1 and w_2, such that w_1 is likely and w_2 is unlikely (like using the foot brake versus driving against a wall). Let us say that $p(w_1)/p(w_2) = 3$ again. There are two actions, a_1 and a_2, which are best responses in w_1 and w_2 respectively. There is only one context. The utilities are given in table 11.

	a_1	a_2
w_1	$5;5$	$0;0$
w_2	$0;0$	$5;5$

Table 11: Example 6

Unlike in the previous example, we assume that there are only two expressions, f and f', which are both unspecific: $[\![f]\!] = [\![f']\!] = \{w_1, w_2\}$. (Or, alternatively, we might assume that f_1 and f_2 are prohibitively expensive.) f' is slightly more expensive than f, like $c(f) = 0$ and $c(f') = 1$.

$$R_0 = \left[\begin{array}{ccc} f/f' & \rightarrow & a_1 \end{array}\right]$$

$$S_0 = \left[\begin{array}{ccc} w_1/w_2 & \rightarrow & f \end{array}\right]$$

$$R_1 = \left[\begin{array}{ccc} f & \rightarrow & a_1 \\ f' & \rightarrow & a_1/a_2 \end{array}\right]$$

$$S_1 = \left[\begin{array}{ccc} w_1 & \rightarrow & f \\ w_2 & \rightarrow & f/f' \end{array}\right]$$

$$\mathbf{R} = R_2 = \left[\begin{array}{ccc} f & \rightarrow & a_1 \\ f' & \rightarrow & a_2 \end{array}\right]$$

$$\mathbf{S} = S_2 = \left[\begin{array}{ccc} w_1 & \rightarrow & f \\ w_2 & \rightarrow & f' \end{array}\right]$$

The crucial point here is that in S_0, the signal f' remains unused. Therefore any rationalizable interpretation of f' which is compatible with its literal meaning is licit in R_1, including the one where f' is associated with w_2 (which triggers the reaction a_2). Robin's reasoning at this stage can be paraphrased as: If Sally uses f, this could mean either w_1 or w_2. Since w_1 is *a priori* more likely, I will choose a_1. There is apparently no good reason for Sally to use f'. If she uses it nevertheless, she must have something in mind which I hadn't thought of. Perhaps she wants to convey that she is actually in w_2.

Sally in turn reasons: If I say f, Robin will take action a_1. If I use f', he may take either action. In w_1 I will thus use f. In w_2 I can play it safe and use f, but I can also take my chances and try f'.

Robin in turn will calculate in R_2: If I hear f, we are in w_1 with a confidence between 75% and 100%. In any event, I should use a_1. The only world where Sally would even consider using f' is w_2. So if I hear f', the posterior probability of w_2 is 100%, and I can safely choose a_2.

If Robin reasons this way, it is absolutely safe for Sally to use f' in w_2.

Example 7 M-implicatures have been used as motivating example for **bidirectional Optimality Theory** (see for instance Blutner 2001) as a framework for formal pragmatics. It has been shown in Jäger (2002) that the set of (weakly) bidirectionally optimal form-meaning pairs can be computed by an iterative procedure that has some similarity to the one given in definition 4. It is thus an interesting questions how the two frameworks relate.[10]

[10] See also Dekker and van Rooy (2000) and Franke (2007) for discussions on how bidirectional OT and game theory relate.

Weak bidirectionality predicts that simple forms are paired with stereotypical meanings and complex forms with atypical meanings. The prediction is even strong though: if the set of forms in question can be ordered according to complexity in a linear way, like $c(f_1) < c(f_2) < \cdots < c(f_n)$, and the set of meanings has the same cardinality and can also be ordered in a linear fashion (like $p(w_1) > p(w_2) > \cdots > p(w_n)$, then the bidirectionally optimal pairs are all pairs (f_i, w_i).

Let us see what pragmatic rationalizability predicts. Suppose there are three worlds with $p(w_1) > p(w_2) > p(w_3)$. Also, there are three forms with $c(f) < c(f') < c(f')$ which are semantically synonymous, namely $[\![f]\!] = [\![f']\!] = [\![f'']\!] = \{w_1, w_2, w_3\}$. There are three actions such that exactly one action is optimal for each world for both players. There is only one context; the utilities are as in table 12.

	a_1	a_2	a_3
w_1	$5;5$	$0;0$	$0;0$
w_2	$0;0$	$5;5$	$0;0$
w_3	$0;0$	$0;0$	$5;5$

Table 12: Example 7

Here is the iterative reasoning sequence:

$$R_0 = \left[\ f/f'/f'' \ \rightarrow \ a_1 \ \right]$$

$$S_0 = \left[\ w_1/w_2/w_3 \ \rightarrow \ f \ \right]$$

$$R_1 = \begin{bmatrix} f & \rightarrow & a_1 \\ f'/f'' & \rightarrow & a_1/a_2/a_3 \end{bmatrix}$$

$$\mathbf{S} = S_1 = \begin{bmatrix} w_1 & \rightarrow & f \\ w_2/w_3 & \rightarrow & f/f'/f'' \end{bmatrix}$$

$$\mathbf{R} = R_2 = \begin{bmatrix} f & \rightarrow & a_1 \\ f'/f'' & \rightarrow & a_2/a_3 \end{bmatrix}$$

Pragmatic rationalizability makes significantly weaker predictions than bidirectional OT. We do predict a division of pragmatic labor in the sense that the cheapest form, f, is specialized to the most probable interpretation w_1 (and the corresponding best action a_1), while the more complex forms f' and f'' are specialized to the non-stereotypical meanings. However, no further specialization between f' and f'' is predicted.

This seems to be in line with the facts. Next to the two expressions in (4), there is a third alternative, which is still more complex than (4b).

(5) John brought the car to a stop.

Also, there are various non-standard ways of making a car stop. The most probable way besides using the foot brake is perhaps to use the hand brake, driving against a wall is even less likely. So bidirectional OT would predict that (4b) carries the implicature that John used the hand brake, while (5) is restricted to even more unusual ways of stopping a car. The present framework only predicts that both (4b) and (5) convey the information that John acted in a somehow non-stereotypical way. While intuitions are not very firm here, it seems to me that the predictions of bidirectional OT might in fact be too strong here.

Example 8 Here is another example that has been analyzed by means of bidirectional OT in the literature. Krifka (2002) observes that the pragmatic interpretation of number words follows an interesting pattern that is reminiscent of Levinson's M-heuristics:

> "RN/RI principle:
>
> a. Short, simple numbers suggest low precision levels.
>
> b. Long, complex numbers suggest high precision levels."
>
> (Krifka 2002:433)

This can be illustrated with the following contrast:

(6) a. The distance is one hundred meter.

b. The distance is one hundred and one meter.

The sentence (6b) suggests a rather precise interpretation (with a slack of at most 50 cm), while (6a) can be more vague. It may perhaps mean something between 90 and 110 meter. Actually, (6a) is pragmatically ambiguous; depending on context, it can be rather precise or rather vague. The crucial observation here is: A shorter number term like "one hundred" allows for a larger degree of vagueness than a more complex term like "one hundred and one."

Krifka also observes that the degree of vagueness of a short term can be reduced by making it more complex — for instance by modifying it with "exactly":

(7) The distance is exactly one hundred meter.

Krifka (2002) accounts for these facts in terms of bidirectional OT, assuming a general preference for vague over precise interpretation. Krifka (2007) contains a revised analysis which employs game theoretic pragmatics. Space does not permit a detailed discussion of Krifka's proposals; in the following I will just briefly sketch how pragmatic rationalizability accounts for Krifka's observations.

Suppose there are two equiprobable worlds, w_1 and w_2. Suppose the distance in question is exactly 100 meter in w_1 and 101 meter in w_2. There are three signals: f_1 (“The distance is one hundred meter.”, f_1' (“The distance is exactly one hundred meter.”) and f_2 (“The distance is one hundred and one meter.”). So we have $\llbracket f_1 \rrbracket = \llbracket f_1' \rrbracket = \{w_1\}$, and $\llbracket f_2 \rrbracket = \{w_2\}$. Let us assume that $c(f_1) = 0$ and $c(f_1') = c(f_2) = 4.5$. There are two actions. a_1 is optimal for w_1 and a_2 for w_2. Furthermore, there are two contexts. In c_1, precision is very important. This means that the differential costs of using an expensive message are lower than the difference in utility between a_1 and a_2. In c_2 it is the other way round.

Table 13 gives the numerical utilities:

		a_1	a_2
c_1 :	w_1	$10;10$	$0;0$
	w_2	$0;0$	$10;10$

		a_1	a_2
c_2 :	w_1	$4;4$	$0;0$
	w_2	$0;0$	$4;4$

Table 13: Example 8

Here is the iterative reasoning sequence:

$$R_0 \quad = \quad \begin{bmatrix} (c_1, f_1) & \rightarrow & a_1 \\ (c_1, f_1') & \rightarrow & a_1 \\ (c_1, f_2) & \rightarrow & a_2 \\ (c_2, f_1) & \rightarrow & a_1 \\ (c_2, f_1') & \rightarrow & a_1 \\ (c_2, f_2) & \rightarrow & a_2 \end{bmatrix}$$

$$S_0 \quad = \quad \begin{bmatrix} (c_1, w_1)/(c_2, w_1)/(c_2, w_2) & \rightarrow & f_1 \\ (c_1, w_2) & \rightarrow & f_2 \end{bmatrix}$$

$$\mathbf{R} = R_1 \quad = \quad \begin{bmatrix} (c_1, f_1) & \rightarrow & a_1/a_2 \\ (c_1, f_1') & \rightarrow & a_1 \\ (c_1, f_2) & \rightarrow & a_2 \\ (c_2, f_1) & \rightarrow & a_1/a_2 \\ (c_2, f_1') & \rightarrow & a_1 \\ (c_2, f_2) & \rightarrow & a_2 \end{bmatrix}$$

$$\mathbf{S} = S_1 \quad = \quad \begin{bmatrix} (c_1, w_1) & \rightarrow & f_1/f_1' \\ (c_1, w_2) & \rightarrow & f_2 \\ (c_2, w_1)/(c_2, w_2) & \rightarrow & f_1 \end{bmatrix}$$

The two complex expressions f_2 and f_1' are alway interpreted in a precise way under the PRSs. The simple expression f_1 is pragmatically ambiguous between a precise interpretation (in c_1) and a vague interpretation (in c_2).

6 Related work

The essential intuition behind the proposal laid out here is that the literal meaning of signals constitutes their default interpretation, and that rational communicators decide about their communicative strategies by iteratively calculating the best response to this default strategy. Similar ideas have been proposed at various places in the literature, sometimes implicitly, even though the precise technical implementation offered here is to my knowledge novel.

As briefly discussed above, Rabin (1990) gives a definition when a message should count as credible. Within the present framework, his definition could be recast as: a message f is credible iff for each n and for each $s \in S_n$, $[\![f]\!] \subseteq s^{-1}(f)$. This equivalence only holds under certain side conditions pertaining to the space of available messages, but essentially Rabin's definition of credibility relies on an iterated calculation of potential best responses, starting with the credulous receiver strategies.

Stalnaker (2005) proposes an informal notion of credibility that could be interpreted as follows: f is credible iff there is some $s \in S_0$ such that $f \in \text{range}(s)$, and for each $s \in S_0 : s^{-1}(f) \subseteq [\![f]\!]$.

Benz and van Rooij (2007) develop a pragmatic interpretation procedure that can, in the present framework, be approximated by the rule: Sally should choose her signals according to S_0, and Robin should interpret them according to R_1. They assume an additional constraint though requiring that only honest strategies will be admitted in S_0.

Both Jäger (2007) and Franke (2008) propose to calculate the pragmatically licit communication strategies by starting with a strategy based on the literal interpretation of signals and iteratively computing the best response strategy until a fixed point is reached. So these approaches are very similar in spirit to the present one. Nevertheless the three theories differ considerably in detail. In Jäger (2007) I assumed a cautious update rule where S_{n+1}/R_{n+1} are mixed strategies that differ only infinitesimally from S_n/R_n. The reasoning process that is modeled this way is quite unlike the Gricean inference schemes that are dealt with in the present framework.

Franke (2008) is conceptually even more similar. The main differences are that Franke uses a particular honest sender strategy — rather than the set of all credulous receiver strategies — as the starting point of the iteration process, and that he uses deterministic best response calculation, rather than potential best responses, as update rule.

7 Conclusion

This article aimed at introducing readers with a background in linguistic semantics and pragmatics to some of the issues that game theorists worry about when study the conditions for communication between rational agents. The question whether or not it is rational to communicate at all in a particular situation has largely been ignored in the linguistic research tradition because a complete alignment of interests is usually assumed. The game theoretic research has shown that communication can be rational even if the interests of the interlocutors are only partially aligned.

A second issue that is prominent in the game theoretic discussion is the role of conventionalized meaning of messages in situations where a simple-minded assumption of honesty and credulity is in partial conflict with rationality. This is also one of the core concerns of Gricean pragmatics. I proposed a game theoretic formalization of Gricean reasoning that both captures the intuitive reasoning patterns that are traditionally assumed in the computation of implicatures, and that addresses the problem of the credibility of signals under partially aligned interests of the interlocutors.

8 Acknowldegments

Thanks to Michael Franke and Robert van Rooij for a very illuminating discussion, and to Michael Franke and Christian Ebert for pointing out several mistakes in a draft version of this article.

References

Benz, Anton and Robert van Rooij. 2007. Optimal assertions and what they implicate. *Topoi - an International Review of Philosophy* 27, 63–78.

Blutner, Reinhard. 2001. Some aspects of optimality in natural language interpretation. *Journal of Semantics* 17, 189–216.

Crawford, Vincent P. and Joel Sobel. 1982. Strategic Information Transmission. *Econometrica* 50, 1431–1451.

Dekker, Paul and Robert van Rooy. 2000. Bi-directional optimality theory: An application of game theory. *Journal of Semantics* 17, 217–242.

Farrell, Joseph. 1993. Meaning and credibility in cheap-talk games. *Games and Economic Behavior* 5, 514–531.

Farrell, Joseph and Matthew Rabin. 1996. Cheap talk. *The Journal of Economic Perspectives* 10, 103–118.

Franke, Michael. 2007. Interpretation of optimal signals, to appear in Krzysztof Apt et al. (eds.) 'Texts in Logic and Games: New Perspectives on Games and Interaction', Amsterdam University Press.

Franke, Michael. 2008. Inference in case of conflict, manuscript, University of Amsterdam.

Grice, Herbert Paul. 1957. Meaning. *Philosophical Review* 66, 377–388.

Grice, Herbert Paul. 1975. Logic and conversation. *Syntax and Semantics 3: Speech Acts*, edited by P. Cole and J. Morgan, 41–58, New York: Academic Press.

Horn, Laurence. 1984. Towards a new taxonomy for pragmatic inference: Q-based and R-based implicatures. *Meaning, Form, and Use in Context*, edited by Deborah Schiffrin, 11–42, Washington: Georgetown University Press.

Jäger, Gerhard. 2002. Some notes on the formal properties of bidirectional Optimality Theory. *Journal of Logic, Language and Information* 11, 427–451.

Jäger, Gerhard. 2007. Game dynamics connects semantics and pragmatics. *Game Theory and Linguistic Meaning*, edited by Ahti-Veikko Pietarinen, 89–102, Elsevier.

Krifka, Manfred. 2002. Be brief and vague! and how bidirectional optimality theory allows for verbosity and precision. *Sounds and Systems. Studies in Structure and Change. A Festschrift for Theo Vennemann*, edited by David Restle and Dietmar Zaefferer, 439–358, Berlin: Mouton de Gruyter.

Krifka, Manfred. 2007. Approximate interpretation of number words: A case for strategic communication. *Cognitive foundations of interpretation*, edited by Gerlof Bouma, Irene Krämer, and Joost Zwarts, 111–126, Amsterdam: Koninklijke Nederlandse Akademie van Wetenschapen.

Levinson, Stephen C. 2000. *Presumptive Meanings.* MIT Press.

Lewis, David. 1969. *Convention.* Cambridge, Mass.: Harvard University Press.

Osborne, Martin J. 2003. *An Introduction to Game Theory.* Oxford: Oxford University Press.

Rabin, Matthew. 1990. Communication between rational agents. *Journal of Economic Theory* 51, 144–170.

Rabin, Matthew. 1992. Corrigendum. *Journal of Economic Theory* 58, 110–111.

Stalnaker, Robert. 2005. Saying and meaning, cheap talk and credibility. *Game Theory and Pragmatics*, edited by Anton Benz, Gerhard Jäger, and Robert van Rooij, 83–100, Palgrave MacMillan.

Zapater, Iñigo. 1997. Credible proposals in communication games. *Journal of Economic Theory* 72, 173–197.