Relevance in Cooperation and Conflict

Michael Franke, Tikitu de Jager, and Robert van Rooij*
Institute for Logic, Language and Computation
University of Amsterdam
{M.Franke,S.T.deJager,R.A.M.vanRooij}@uva.nl

August 29, 2008

Contents

[*] Author names appear in alphabetical order. We would like to thank . . .

1 Introduction

The enterprise of Gricean pragmatics can be summarised as exploring the inferences (beyond the semantic content of an utterance) licensed by the assumption that a speaker intends to contribute cooperatively to some shared conversational end, and the extent to which these inferences match those which real hearers and speakers respectively draw and intend their listeners to draw (Grice, 1989). The Gricean maxim "Be relevant" is a classic case in point: if a speaker flouts the maxim, uttering an apparent *non sequitur*, the assumption that the speaker is nonetheless cooperative leads a hearer to put extra effort into discovering what —beyond the apparently irrelevant semantic meaning— the speaker could have intended to convey.

There is an inherent self-referential loop in this description: a speaker must have expectations about a hearer's interpretative response (the additional inferences he might be capable of imagining or willing to draw), while the hearer must tailor his response to his notion of these speaker expectations, and so on into deeper nestings of belief operators.

Classical game theory studies just such situations of interlocked multi-agent belief and action (where "action" here represents both the utterance itself and the interpretation that a hearer eventually must form). That is, when several agents must independantly choose how to behave in a situation whose outcome depends on their collective behaviour, the tools of game theory are *prima facie* appropriate.

However classical game theory is not typically cooperative; it deals rather with situations of at least partial conflict, where not all agents prefer the same possible outcomes. The idea of agency in game-theory is that of a self-interested *homo economicus*. If speaker and hearer preferences do not coincide, then they may not agree on which contributions to the conversation are relevant; we can distinguish speaker relevance from hearer relevance, meaning relevance to the individual goals of the conversational participants. The general idea of this paper is that *speaker* relevance plays a significant, and so far largely overlooked, role in pragmatic reasoning.

The idea of "speaker-relevance" lets us extend the scope of linguistic pragmatics beyond cases of pure cooperation. Recall that Gricean pragmatics is deeply rooted in the Cooperative Principle: the notion that a (normal) conversation has a common purpose or direction, shared by all participants. But natural examples of speaker-hearer conflict are not hard to find, from lying and keeping secrets in a court of law through to avoiding a sensitive topic with dinner guests or feigning interest in a boring story badly told. One of the main concerns of this paper is to argue that such cases where speaker and hearer interest diverge —partially or wholly— should be incorporated into the pragmatic agenda, for they can be studied no less systematically than their pure cooperative cousins. The claim to be defended throughout is that game theory is the tool of choice to study the impact of cooperation and (partial) conflict in conversation, because it allows for fine-grained distinctions in modelling agents' preferences (and knowledge thereof).

The question arises how the idea of an entirely self-interested speaker is compatible with the obvious factual prevalence of cooperative linguistic behavior. In Section 2, we seek to reconcile this apparent paradox by arguing towards the idea that cooperation can be expected *in the majority of cases* in a society where providing truthful and hearer-relevant information pays for the speaker in terms of *social status*. In other words, while speakers are indeed self-interested and all speaker behavior is first and foremost "speaker-relevant", there may be external pressures —like the threat of punishment or status loss— that align or quasi-align interlocutors' preferences.

If this is true then cooperation is a phenomenon that can often, but not always, be expected. The rest of the paper concerns the kinds of pragmatic inference that can be made in cases of *partial* conflict: preferences that diverge enough to allow for interesting strategic reasoning, but not so widely that no communication at all is possible.

More concretely, here are some examples of the kind that we will be concerned with. Suppose two people are arguing about the conflict between Israel and Palestine; one, claiming that the blame lies with Palastine, says:[1]

(1) Most Israelis voted for peace.

The argumentative setting clearly distinguishes this example from more standard cases of scalar implicature. Intuitively the inference that not all Israelis voted for peace is one the speaker *does not want* the hearer to make, but still it seems entirely justified. We will investigate this example further in Section 3 in the context of PERSUASION GAMES.

Another example of pragmatic inference beyond pure cooperativity is what we will call the 'unwilling to tell'-inference:

(2) A: Where are you going?
B: Out.

Unlike the previous example, this may well be an inference that the speaker wants the hearer to draw. But nevertheless, as we will discuss in Section 4, most contemporary pragmatic theories which rely on full speaker-hearer cooperation have difficulties explaining this inference.

Section 4 also covers the interesting case of attempted deception without lying, by unviting pragmatic inferences that would only be appropriate if the speaker were fully cooperative. The following exchange from the Clinton impeachment proceedings (taken from Solan and Tiersma (2005)) exemplifies the point:

[1] The example stems from a bumper sticker and is discussed at length in Ariel (2004). No inferences regarding the political opinions of the authors should be drawn.

(3) Q: At any time were you and Monica Lewinsky together
alone in the Oval Office?
A: [I]t seems to me she brought things to me once or twice
on the weekends.

In Section 5 we argue that recognising deception should also be considered pragmatic reasoning, and show the difficulties contemporary theories of pragmatic inference have with this extension.

Section 6 extends the argument to cases of flat-out lying, where still it is pragmatic reasoning that allows a hearer to decide whether the speaker's utterance should be taken as credible or not.

The key suggestion throughout is that it is an explicit representation of the speaker's preferences that enables the fine-grained distinctions we want to make: between truly cooperative speakers and those with their own agendas, and between different kinds of pragmatic inference hearers may make.

2 Cooperation and conflict

Taking a game-theoretic view of language use requires a formal model of linguistic interaction as a game. The standard model that has emerged for applying game theory to questions of pragmatics is the SIGNALLING GAME (the form was first introduced in Lewis (1969); applications to pragmatics can be found in Parikh (2001) and Benz et al. (2006)). We will sketch the essential features of the model here, deferring the technical details until Section 3.

A signalling game is played between a Speaker and a Hearer. The speaker has some information that the hearer lacks, but which he would like to have in order to decide on some action to take. The speaker sends a message to the hearer, which he can use to condition his action. The payoffs to both speaker and hearer depend in general on the private information the speaker holds (often thought of as some 'state of the world' which she has observed but which the hearer has not), the message, and the action. Typically different states should induce different best actions from the hearer, according to his payoff function; a standard interpretation is that such an action is simply forming a belief about the state, so that each state corresponds to exactly one best action (believing that that particular state obtains), and vice versa.

Gricean cooperation is typically modelled in this setting by giving speakers and hearers the same payoff function. This ensures that whatever is in the one's interests is in the interests of the other, and thus that both prefer cooperation. Once the payoffs diverge, however, we can have situations of (partial) conflict: the speaker, knowing the state, may prefer the hearer to take an action that differs from his best choice according to his own preferences.

It is easy to see that under conditions of extreme conflict (a zero-sum game), no communication can be sustained. For why should we give information to the enemy, or believe what the enemy tells us? Communication can only succeed if preferences are aligned enough, enough of the time, to sustain it. In this section we will argue that this alignment of preferences is surprising from the game-theoretic point of view, and requires some additional explanation. Put more strongly, the non-cooperative game-theoretic view of language has difficulty explaining the apparent ubiquity of hearer-centred relevance: speakers who apparently subordinate their own interests to the interests of their interlocutors.

We can highlight the problem by considering the origins of human speech. Of course this can only be speculation, but the difficulty is most apparent when we imagine a society in which failure to coordinate on communication will not immediately result in incarceration; a society, that is, in which there is still a real social alternative to cooperative language use.

2.1 Evolutionary origins and altruism

We begin with an apparent paradox. If speakers and hearers are considered as evolutionary agents, in competition for the scarce resources required for reproduction, the mere transfer of information from speaker to hearer immediately seems to give rise to opposing preferences; this, in turn, seems to preclude informative language use. Put differently, the non-cooperative game-theoretic view of language use faces the significant challenge that it seems to predict absolutely no informative language use whatsoever.

To make the problem clear we must first justify the notion that information transfer is an altruistic act. Clearly the information received is valuable to the hearer (this is directly encoded in the payoff structure of the game). The natural interpretation of this structure is that the information truly reflects some state of the world, and it is this state which the action of the hearer should match. But in that case it seems that the speaker would be giving away a potential advantage by communicating her information to the hearer; she loses the opportunity to exploit her information without competition at some later date.

Even when this is not the case (when, for whatever reason, the speaker is permanently barred from making use of her information) merely helping a potential competitor must be considered an altruistic act. It is most plausible that language arose in a species living in relatively small groups with a strong social structure. A high degree of cooperation can be expected in such societies (partly based on kin selection and partly on reciprocity and simultaneous collective action enforced by explicit social pressure), but informative discourse seems a perfect candidate for deviations from cooperation.

In small tightly-knit social groups, social competition must be much more restrained than for species that form large herds or that do not band together at all. A small group relies on each individual member far more than a large herd, so competition between group members can best alter only their relative standing within the group without significantly damaging their ability to contribute to group functioning. If I fail to tell you that there is ripe fruit over here where I'm sitting I may gain a small fitness advantage over you; if I fail to tell you there is a poisonous snake where you're about to sit and you are bitten, the loss to the group far outweighs my gain in reduced competition.

It seems that the complex structure of human language, with its potential for representing closely-related meanings, is well suited to competition based on fine distinctions; the animal signal systems typically considered most similar to human language, on the other hand, tend to be used in extreme cases analogous to the snake in the grass (the most famous example is probably that of vervet monkey alarm signals, see Cheney and Seyfarth (1990)). But if human language is uniquely well-suited to relatively harmless competition between members of the same social group, how does cooperative language use arise?

Altruism in general is an interesting problem for evolutionary biology, with a number of different models appropriate for different species and settings (Hamilton, 1963; Trivers, 1971; McAndrew, 2002). The most promising for the current setting is reciprocity (Trivers, 1971): the idea is that an altruistic act might be performed in the expectation that the recipient will repay the benefactor at some later date. However this explanation in its basic form makes predictions that are not borne out in the case of language. Most obviously, speakers should only dispense their information when they expect the hearer to reciprocate in kind. More subtly, an asymmetry in 'cheater detection' is predicted which does not seem to match actual speaker/hearer behaviour.

The problem with reciprocity as a strategy is that it is vulnerable to invasion by 'cheaters': individuals who reap the benefit of others' altruistic acts but never get around to performing any themselves. Standard models show that making a system of reciprocity cheater-proof requires two things: cheater detection and punishment. In the linguistic case, speakers should keep track of the hearers they have given information to, and monitor whether they do in fact reciprocate the favour (cheater detection); and they should punish those who do not. (See for example Bowles and Gintis (2004).)

The first of these predictions already seems odd, and the second is clearly nonsense. Zahavi and Zahavi (1997) have shown that many kinds of animal signals do not display the properties we would expect if information-giving is taken to be equivalent to utility transfer, and Dessalles (1998) has extended the argument to human language. People are typically *eager* to speak, if they have information that they think might be useful. We do not castigate the humble listener who politely attends our discourse and adds only encouraging nods and carry-ons, but the bore who spins his tale out interminably so that we 'can't get a word in edgewise' (indeed, to be 'a good listener' is a rare and valuable skill). Speakers do not watch hearers for cheating; on the contrary, *hearers* check *speakers* for accuracy and (most interestingly) for the genuine newness of their information. Speakers even *compete* for the privilege of adding their unique information to a conversation. This suggests another way out of the altruism problem. Perhaps the information being transferred is not the only utility-bearing element of the transaction: speakers, too, are getting something directly out of their participation in conversation.

2.2 Status as realignment mechanism

Dessalles suggests that this something is STATUS. The notion can be traced back to Zahavi (1990) in investigations into the behaviour of Arabian Babblers (a bird which lives in small but highly structured social groups). Status is a social parameter reflecting influence within a social group; crucially, it is conferred by the group rather than claimed by the individual. The idea is that advice-giving confers status on the giver, while advice-taking reduces it for the taker; the exchange of information for status is reasonable whenever the value of the information for the hearer is sufficiently larger than its value for the speaker (in terms of the immediate uses the information may be put to) to justify the exchange in the eyes of both parties.

If status were, like information, a commodity that hearers themselves could decide to confer or withhold, this account would suffer from just the same problems as the reciprocity story: a hearer would have an incentive to refuse to confer status even after receiving good advice, and we would again need cheater-detection by speakers and so on. However the fact that status is conferred by the group at large avoids this problem. Babblers accrue status from potentially dangerous activities such as standing guard, and from giving food to other babblers (both apparently altruistic acts); the birds perform these actions ostentatiously, making sure they are observed by the other group members whose approval they are courting. "Didn't I warn him?" is the victorious cry of an advice-giver claiming the status she is due.

This account also makes predictions about speaker and hearer behaviour, but this time the predictions are largely borne out by observation. Speakers should compete to provide information, in the hopes of the status improvement it will bring them; hearers will need to scrutinise this information for plausibility, since speakers have an incentive to lie if they can get away with it (speakers who monopolise their status-gaining position without conveying useful information, even if they do not actively lie, will also be censured).

The non-cooperative game-theoretic view of language makes the hearer-centered notion of relevance entirely inexplicable. Status goes some way towards remedying this, by showing the speaker's incentive to make altruistic gifts of information in order to accrue status in the eyes of observers. One more element is needed to complete the picture: hearer choice.

Unlike in the one-shot game-theoretic setting, a hearer (or advice-seeker) is not restricted to listening to a single interlocutor. We already mentioned speaker competition, which brings with it the possibility for hearers to actively select the speakers they wish to be informed by. Clearly it is in the hearer's interests to select the speaker they believe most likely to assist them according to their current needs: a speaker with expertise in whatever their current difficulty is, and one who is willing to exercise that expertise in their interest.

How might a speaker advertise her ability and willingness to apply it? She could easily claim to be expert in bicycle repair (if that is the matter at hand), but the same claim can just as easily be made by all her competitors, whether true or not (we will take up the question of credibility in more detail in Section 6). Much better is to offer a piece of specific advice that is both relevant and helpful; this gift of information will result in her own status gain, but more importantly it acts as an *advertisement* of her abilities which (if all goes well) will lead to the hearer continuing to seek her advice in the future.

If every communication from speaker to hearer is seen in this dual role, as both information-giving and advertisement, the importance of relevance becomes clear. A speaker expert in 19th-century philosophy can offer information to a hearer with a flat tyre, but the hearer can better reject this offer and seek advice elsewhere. (Such a speaker, in turn, is better off searching for a philosophy student in need; differentiated expertise and hearer choice leads to what we could call 'assortative mixing' where expert speakers become matched to hearers with interests in precisely their field of expertise.) If a hearer expects the speaker to value the continuation of their conversation (as a status model predicts), he is justified in the assumption of relevance: a speaker will know that the hearer's attention is limited and will strive to monopolise it by making his advice helpful and to the point.

In fact much of the Gricean apparatus can be supported in this way. So long as hearers can exercise choice, a speaker who wishes to continue the conversation is required to adjust her efforts to (her understanding of) the hearer's needs; typically this requires her to be truthful (and we seem also to cover a range of fictional settings without difficulty), to express herself concisely and clearly, and to stick to the point. Doing otherwise will quickly lead to the hearer turning his attention elsewhere.

In other words, by letting the hearer select his advice-giver and a speaker gain status from giving advice, we have given a reason to expect (largely) aligned payoffs in normal situations. A speaker has an incentive to do *whatever will keep the hearer happy*, as this is what will lead to her own status gains. Clearly this incentive is subordinate to particular gains she might earn by concealing information she can use directly herself, so that in clearcut cases of conflict of interest we expect the status-based Gricean maxims to be relaxed; and indeed if we observe extreme cases such as courtroom testimony this seems to be the case (see below, in particular section 4). In short, this perspective justifies *largely* aligned preferences *most* of the time, while leaving open the possibility of preference conflicts that the following sections will explore.

Taken together, the game-theoretic perspective suggests that speakers are after all self-interested. Hearer-relevant behavior emerges when external motivations align the speaker's preferences sufficiently with the hearer's.[2] It is certainly interesting to dwell on the exact mechanism (perhaps automatic) with which speaker preferences adjust to the hearer's concerns.[3] But we will not do so in the remainder of the paper. Rather we will look at a variety of cases where speaker preferences do not align entirely with the hearer's. (We will not

[2] The economic and biological literature is rich with studies of how and under which circumstances *costly signaling* can enforce honest and revealing, i.e. hearer relevant, signaling. The most influential early work is Spence's (1973) analysis of job market signaling and Zahavi's (1975) and Grafen's (1990) signaling game models of mate selection. Costs in these models can be looked at as a blackbox-parameter for whatever external mechanism influences the speaker's strategic decisions on top of his primary material concerns.

[3] Natural candidates are ethical and moral considerations, as well as patterns of learned social behaviour. Seminal work on 'psychological game theory', incorporating these non-material payoffs in the utility function, is due to Geanakoplos et al. (1989) and Rabin (1993).

be concerned with why that is.) Throughout Sections 3, 4 and 5 we maintain the assumption, which is fairly central in linguistic pragmatics, that speakers speak truthfully. In section 6 even this assumption is dropped, and we consider the most general case of pragmatic inference under conflict.

3 Persuasion with verifiable messages

Once we have a picture of hearers policing speakers for cooperative behaviour, a very natural extension is to assume that messages are verifiable. The idea is simply that speakers will be punished for telling outright lies, although they may misdirect or be deceptive in more subtle ways (compare courtroom testimony, where a witness can be convicted for perjury but need not volunteer information; see section 4 in particular). We will see that a range of implicature-like inferences can be drawn in this setting, despite the presence of conflicting interests. In particular, in this section we will discuss examples like (1) in which the speaker, in some sense, does not want to communicate a certain pragmatic inference, but cannot prevent that it is nevertheless drawn. To make the discussion precise, however, we will need the formal details of the signalling games that were introduced in broad outline in the previous section.

3.1 Signalling games

Recall that a signalling game runs schematically as follows: a Speaker has some private information which a Hearer lacks; the Hearer must make a decision which he would like to condition on this information. The Speaker sends a signal, and the Hearer may condition his decision on this signal; the aim of a game-theoretic analysis is to isolate strategies of signal sending and interpretation that model standard cases of pragmatic reasoning.

Formally, we model the information held by the speaker as a random move by Nature. Speaker sees only Nature's move (a 'state of the world', also known as a SPEAKER TYPE) and produces a message; Hearer sees only Speaker's move (the message) and chooses an action (which we will typically call an INTERPRETATION). The payoff for speaker and hearer depends in general on all components: state, message and interpretation.

Let T be a set of 'states of the world' or speaker types, with typical element t (sometimes it will be useful to think of these as how the world really is, other times as information or dispositions the speaker herself might hold). M is the set of messages at the disposal of the speaker, and for any $m \in M$ the SEMANTIC MEANING of m, written $[\![m]\!]$, is some subset of T. Finally, A is the set of hearer actions; it will often make sense to set $A = T$, with the interpretation that the hearer is trying to discover from the sender's message what information she holds.

To model the information restrictions we use the notion of a STRATEGY. A strategy for some player is a plan dictating the choices that player will make in every situation she might find herself in. A (pure) strategy for a speaker is a function $\sigma\colon T \to M$, which decides for each state she may find herself in what message she will send. A pure hearer strategy is likewise a function $\rho\colon M \to A$ giving the action the hearer takes on receipt of each message. A MIXED strategy is a probability distribution over pure strategies, or (equivalently in this setting) a function from states to distributions over messages (for the speaker) or from messages to distributions over actions (for the hearer). We may talk also about Nature's 'strategy', which is a degenerate mixed strategy: it is simply a probability distribution over states.

Assuming Nature's strategy puts positive probability mass on all states, we say a pair of strategies for speaker and hearer are in NASH EQUILIBRIUM (or just "equilibrium") when playing them ensures that neither player has an incentive to change strategies. As we will see later, not all equilibria are equally convincing as models of linguistic behaviour, but non-equilibrium combinations of strategies should certainly be ruled out.

Eventually we will consider in full generality information transfer given non-aligned preferences. We begin, however, with the special case of verifiable messages.

3.2 Persuasion games

PERSUASION GAMES are models of buyer/seller interaction, originating in the economics literature, in which the message the seller sends is VERIFIABLE (more or less equivalently, in which the seller is required to tell the truth) (Milgrom and Roberts, 1986).

We model this simply by a requirement on speaker strategies: we consider only strategies σ that meet the requirement that for all states $t \in T$, $t \in [\![\sigma(t)]\!]$. That is, if the speaker says m in state t, then t should be part of the semantic meaning of m (this restriction is generally thought of in the game-theoretical pragmatics literature as an implementation of Grice's maxim of Quality). It is equivalent to imagine that there are penalties for sending false messages and a sufficient chance of being caught (which perhaps gives a closer match to intuitions about the ordinary case), so long as the penalties are large enough that speakers never have an incentive to lie.

While it is the presence of verifiable information that defines a persuasion game, typically the models in the economics literature (and those we are concerned with here) have some additional structure. In particular, the preferences of the speaker over hearer actions are independant of the actual state of the world (that is, the payoff function for the speaker ignores the state of the world, and in fact the message sent as well). These preferences induce a linear order on hearer actions according to the preferences of the speaker. (The case is similar to standard scalar implicature, except that the preferences of speaker and hearer diverge so the inference cannot be based on assumptions of cooperation.)

We will assume, following the economics literature, that for each state t there is a unique corresponding action a_t which the hearer prefers when t is the case. Since we think of these actions as interpretations, we can unify the sets T and A and say that the hearer prefers to believe t iff t is in fact the case.

For convenience we number the states according to the speaker's preference $t_1, \ldots, t_n$, where t_n is the most preferred interpretation and t_1 the least. (This ordering is on interpretations; the speaker need not prefer state t_2 over t_1, but does prefer that the hearer believe that t_2 obtains rather than t_1.)

Intuitively, then, the game looks like this: Nature provides the speaker with a state; the speaker sends a true (but not necessarily complete) message about which state obtains; finally the hearer forms an interpretation (chooses which state he believes obtains). The buyer/seller interaction, although not explicitly modelled, gives a good intuition for the agents' respective preferences. If the state represents the true quality of some product, the seller prefers the buyer to believe that the quality is high regardless of the truth, while the buyer would like to know what is actually the case.

Milgrom and Roberts show conditions under which all equilibria in the persuasion game are fully revealing, in the sense that the hearer learns the true state despite the speaker's interest in concealing it. The proof takes the form of an UNRAVELLING ARGUMENT, an induction down the scale. When the true state is t_n the seller wants the buyer to know it; the message with meaning $\{t_n\}$ may only be sent in this state so the hearer can safely believe it. When the state is t_{n-1} the seller would like the buyer to believe that it is t_n but she may not lie; the best she can do is announce $\{t_{n-1}, t_n\}$, but the buyer will correctly conclude that the state is *not* t_n since if it were the seller would have sent the more specific message.

This schematic presentation is intuitive, but the devil is in the details. The unravelling argument does not, in fact, succeed for the games we are considering, where hearer actions are interpretations and thus discrete objects. The theorem that Milgrom and Roberts prove requires that hearer actions lie on a *continuum* (they correspond to quantities of the good purchased, rather than direct beliefs about quality) and this assumption is required for the proof. To see that the argument fails for pragmatic interpretation games, let us consider the simplest non-trivial scale, with two elements, and the "seduction game".[4]

In (the motivation for) this game the speaker wants to seduce the hearer; the hearer, on the other hand, wants only to be seduced by a speaker of high quality. That is, the hearer's interest is in learning the true quality of the speaker, while the speaker's interest is in being taken to be of high quality. We give the speaker two messages: "I am of high quality" and "I am at least of low quality" (this second message is strictly uninformative if only low and high quality are possible, so we gloss it as "Anything at all").

The notation we introduce here will be used for all games in the rest of the paper. Each row corresponds to a state that Nature might choose (we assume that Nature chooses states with equal probability unless stated otherwise); hearer actions are columns in the left portion of the grid; the sender's payoff are given first (if payoffs also depend on which message is sent then we need to enrich the representation). Whenever the semantic meaning of messages is

[4] The name acknowledges the origin of the game in evolutionary biology, as a representation of mate selection. It has been used to motivate costly signalling theory, see footnote 2 on pg. 8.

important, we give it in the right portion of the grid: a tick indicates that the message of that column is compatible with (may truthfully be used in) the state of that row.

	$\texttt{high}_a$	$\texttt{low}_a$	$\texttt{high}_m$	$\texttt{anything}_m$
$\texttt{high}_t$	$1, 1$	$0, 0$	$\surd$	$\surd$
$\texttt{low}_t$	$1, 0$	$0, 1$	$-$	$\surd$

Suppose now that according to Nature's distribution, the type $\texttt{high}_t$ is more likely than $\texttt{low}_t$. It is easy to see that the following strategies are in equilibrium: $\sigma(\texttt{high}_t) = \sigma(\texttt{low}_t) = \texttt{anything}_m$ (the speaker is always uninformative, regardless of the state), and $\rho(\texttt{high}_m) = \rho(\texttt{anything}_m) = \texttt{high}_a$ (the hearer always believes the speaker is of high quality, regardless of the message). This is a classic POOLING EQUILIBRIUM, in which no information is transmitted because all the sender types send the same message. The sender has no incentive to change strategies since the hearer is already taking the action she most prefers; the hearer is maximising expected utility given Nature's distribution and is getting no information from the message so he also has no incentive to change his behaviour.[5]

The difference between this example and the continuous case is the following: in the setting of Milgrom and Roberts (1986), with a continuous space of hearer actions, the hearer's best response to $\texttt{anything}_m$ lies in the *interior* of the interval $[\texttt{low}_a, \texttt{high}_a]$; in the economic setting these actions are real numbers, representing quantities of the good to be ordered. In that case the speaker *would* do better to announce $\texttt{high}_m$ when this is true, and the unravelling argument can proceed. Since we take interpretation actions as discrete, we must repair the situation in a different manner.

Naivety unravelling We introduce three additional assumptions: (1) we order the messages along an entailment scale that matches the speaker's preferences, (2) hearer interpretations respect semantic meaning, and (3) there is an ϵ-probability of 'naive semantic interpretation' by the hearer.

To begin with the first restriction: Recall that we named the states $t_1, \ldots, t_n$ in order of increasing speaker preference (that is, the speaker prefers that the hearer believe the state is t_2 over t_1, t_3 over t_2, and so on, with t_n being the most preferred interpretation). We restrict the messages to the set $\{m_i \; ; \; 1 \leq i \leq n\}$, where $[\![m_i]\!] = \{t_i, \ldots, t_n\}$. Intuitively m_i means "At least t_i", with the proviso that the set of alternatives be finite. (This condition actually held for the seduction game given above, but not in general for the games economists are interested in.)

The second assumption looks unnecessary on the face of it, since hearer best responses should make use of the restriction of speakers to true statements. However we need also to consider the hearer response to *unsent* messages, since

[5] It should be noted that the pooling equilibrium is not a strict equilibrium: the sender is equally well off sending $\texttt{high}_m$ in state $\texttt{high}_t$. The pooling equilibrium is then not evolutionary stable. However, we are here concerned with an argument that gets rid of the pooling equilibrium as a viable solution of one-off rational deliberation.

these can make equilibria non-strict and interfere with the reasoning. (In the pooling equilibrium for the seduction game the hearer strategy might just as well respond to a new message "I am of low type" by believing the speaker to be of *high* type; since the message is never sent, the hearer receives no penalty for this perverse interpretation strategy. We could rule out such oddities with a solution concept such as sequential equilibrium; for simplicity we leave it as a stipulation.)

The naivety assumption is more subtle. In each round of the game, with some small probability ϵ the hearer's response is *not* governed by his strategy (which represents the outcome of pragmatic and strategic reasoning) but by the following rule: he chooses with uniform probability one state from the semantic meaning of the message. The intuition is the following: occasionally a hearer is tired, or distracted, and doesn't perform any strategic reasoning at all. He simply takes the message at face value, according to its semantic meaning; since we don't allow multi-state interpretations, we have him choose one state from the meaning with uniform probability.

According to a population interpretation of the game, in which each round is played with speaker and hearer picked at random from a large population, an ϵ-proportion of the hearer population are agents of extremely bounded rationality: so restricted that they perform no strategic reasoning at all (they are 'naive' or 'credulous' interpreters) and are even blind to the distribution on states given by Nature. We will see in Sections 4.2 and 6 how other levels of strategic sophistication between this extreme naivety and full-blown hyperrationality can be of interest for pragmatic reasoning.

The effect of the naivety assumption can be seen in the seduction game: when the speaker is of high type, if he sends the message "high or low" ($\texttt{anything}_m$) he suffers a small penalty compared to the message "high" *even when the hearer interprets this message the way he prefers.* This is because with probability ϵ the hearer interprets naively; that means that an interpretation of "$\texttt{low}_m$" is possible (with probability $\frac{1}{2}\epsilon$) if the message is "high or low" but impossible if it is "high". Just as in the economic case, then, the speaker strictly benefits by being more specific when she is of high type, and the counterexample is no longer an equilibrium.[6]

The unravelling argument Armed with these assumptions we can now give the unravelling argument for our setting (NAIVETY UNRAVELLING). We proceed by induction. For convenience we'll talk about states and messages being ordered by their indices: $m_i < m_j$ if $i < j$, and so on.

First consider the messages the sender might use if she is in state t_n. Any message is truthful, so we should consider the hearer response to all messages. By assumption (2), $\rho(m_n) = t_n$ (the unambiguous message announcing the true type of the speaker is interpreted correctly). Now consider any other message

[6] Note that this is *not* trembling hand perfection Selten (1975), although the motivation in terms of the possibility of error is similar. The errors the hearer can make are limited by the semantic meaning of the messages, and we do not need to take limits under diminishing error probabilities.

$m_k < m_n$. If $\rho(m_k) \neq t_n$ then the speaker would prefer to use m_n (since she strictly prefers t_n over all other interpretations). Suppose then that $\rho(m_k) = t_n$; still, by the third assumption of hearer error, she prefers to send the unambiguous message.

To see this we only need to look at the possibilities for hearer error: with m_n there is no such possibility, while since $k < n$ there is at least one state t_k such that $t_k \notin [\![m_n]\!]$; with probability ϵ the hearer interprets the message according to its semantics and may make an error interpreting m_k (and one which the speaker prefers him *not* to make) which is not possible with m_n.

So we have established both that $\rho(m_n) = t_n$ and that $\sigma(t_n) = m_n$. Now we consider the inductive step. Suppose that for some i such that $1 < i < n$, for all $j > i$ (with $j \leq n$), $\sigma(t_j) = m_j$ and $\rho(m_j) = t_j$. That is, suppose that the strategies are in perfectly revealing equilibrium for the top of the scale, down to some element i.

If the message m_i is actually sent, it's easy to see that the hearer's best-response interpretation is $\rho(m_i) = t_i$: the sender may not use the message in any lower state (it would be untrue) and will not in any higher (by the inductive hypothesis). We only need to show that the message *will* be sent and the proof will be complete.

Suppose then, towards a contradiction, that at equilibrium the message m_i is not used. By the second assumption, nevertheless we have that $\rho(m_i) \geq t_i$ (hearer interpretations respect semantics, even for unsent messages). Since the speaker prefers (perhaps non-strictly) not to use m_i we have $\rho(\sigma(t_i)) \geq \rho(m_i)$; let us say $\sigma(t_i) = m_k$, then $\rho(m_k) \geq \rho(m_i) \geq t_i$.

Suppose one of these inequalities is strict: then $\rho(m_k) > t_i$. But then the hearer is not playing a best response: m_k is never sent in a state higher than t_i (these were already taken according to the inductive hypothesis and $m_k < m_i$) so he would earn better payoff by changing his interpretation.

But if neither of the inequalities is strict (that is, $\rho(m_k) = \rho(m_i) = t_i$) then the speaker earns more by using m_i, according to the ϵ-probability of error: the hearer interpretation is the same, but $[\![m_k]\!] \supset [\![m_i]\!]$ (because $m_k < m_i$) so the extra possibility of error reduces the speaker's payoff.

This provides the contradiction we needed: the message m_i must in fact be used, and used only in (and interpreted as) t_i. This completes the inductive step, and in turn the proof: the fully revealing profile is the only equilibrium satisfying the conditions we have stated.

The argument transfers quite directly to hearer reasoning, as we can see in the seduction example introduced above. On hearing "High or low" the hearer reasons as follows: "If she were of high type she could say so; indeed she would *rather* say so because that way there's no chance of me not understanding (so she gets her preferred outcome with certainty). Then since she doesn't say so this isn't the case; the only alternative is that she is of low type."

3.3 Pragmatic inferences unintended by the speaker

Classical scalar implicature arises in a cooperative setting; hearers can assume that more specific information would have been given if this were possible, because their interest in accuracy is reflected in the speaker's payoff. The unravelling argument shows that the same pattern of inference can arise in cases of conflict, if the speaker's payoffs induce the right kind of preference for specificity. The difference shows in a rather subtle distinction in how we should characterise the additional inference, normally called an implicature.

In the cooperative case, a sentence like "Most of our office workers play badminton" would be taken to mean "Most (but not all) of our office workers play badminton". The standard reasoning is that the hearer is interested in precise information; the speaker could have said "all", which if true would have been more precise; since she prefers whatever the hearer prefers, she would have done so had she been able to. The final step in the reasoning is to assume that the speaker is expert in the matter at hand: she knows whether all or most-but-not-all of the office workers in fact play badminton. Then if she couldn't make the stronger claim, this is because it is false, and the implicature strengthening the proposition is justified.

According to this story, the implicature is an inference that the speaker wants the hearer to draw; indeed, in situations where this seems not to be the case the inference is not drawn. ("Oops, I used arsenic instead of sugar in the muffins!" "Oh no, I ate some, call a doctor!") The unravelling argument for persuasion games, however, shows that (an inference similar to) the implicature sometimes *is* drawn despite the speaker not intending that it be conveyed.

Persuasion games let us account for example (1) discussed in the introduction. Recall the example and the context: in an argument about the conflict between Israel and Palestine, the speaker, in an attempt to argue that the blame lies with Palastine, says "Most Israelis voted for peace." Intuitively, the scalar inference that (according to the speaker) not all Israelis voted for peace is an inference that does come about even though the speaker does not want to convey it explicitly. The inference goes against her interests, but, being forced to speak truthfully in a reasonable debate, she cannot prevent this inference being drawn. The situation concerning the interlocutor's preferences is exactly the one modelled in a persuasion game and the reasoning required to establish the scalar inference is nothing more than the unravelling argument: the speaker would have preferred to make the stronger claim if possible, but since it isn't true she must content herself with the weaker.

Whether this inference should be called an "implicature" depends on the definitional stance one takes; if implicatures are part of what the speaker *intends* to convey, then this isn't an implicature. We are not concerned with the definitional question: these are inferences, beyond the semantic meaning of a message, that a hearer may draw based on considerations of strategic language use, which at least places them squarely in the pragmatic camp.[7]

[7] The conviction that an implicature must be speaker-intended is widely held, albeit often fairly implicitly. However, Ariel (2004) argues, based on this conviction, that the example in

In a similar example, partial conflict can give rise to 'partial exhaustive' interpretations of wh-phrases. Suppose I am being interviewed for a job in a company based in Portugal. We've discussed the fact that the position is 'live-in': if I take it I will move there, along with my wife and our two children. During the interview, the following exchange takes place (Hirschberg, 1985):

(4) A: Do you speak Portuguese?
B: My wife does.

There are standardly two ways to interpret a wh-phrase as the answer to a question: either EXHAUSTIFIED ("My wife does and nobody else") or as a MENTION-SOME answer ("My wife is someone who does, and anybody else might as well") (Groenendijk and Stokhof, 1984). In this setting, if the salient domain is taken to be our family, neither of these seems correct. The inference we would expect to draw is that I myself do not speak Portuguese, but that either of our children might — a sort of mixture of mention-some and exhaustification.[8]

The situation is not precisely analogous to the scalar case, since the possible alternative answers of the form "X speaks Portuguese", where X ranges over the speaker's family members, are not naturally ordered by entailment. However there seems to be a clear salience ordering, in which my own linguistic ability is most relevant, followed by that of my wife (who presumably might be involved in my work), and then by our children. As in the purely scalar case, I would like to make the strongest claim I can, but truthfulness prevents me from claiming outright a linguistic ability I do not hold. The inference that I do not speak Portuguese is one I do *not* want the hiring committee to draw, but which they certainly will; on the other hand, I am indifferent about their beliefs regarding our children and so they will draw no conclusions.[9]

4 Further pragmatic inferences beyond cooperation

The previous section showed how scalar inferences can arise in situations where the preferences of interlocutors diverge substantially — substantially enough to say that the speaker does not want the inference to be explicitly conveyed, but not enough so as to prevent the inference entirely or even cause a breakdown

(1) can therefore *not* have the (scalar) implicature "not all", and she suggests that the "not all"-component is part of the semantic meaning of '*most*'. In contrast, Sperber and Wilson (1995) explicitly include unintended inferences under the label 'implicature': "Sometimes, the addressee may justifiably attribute to the communicator an implicature that she never in fact intended to communicate" (Sperber and Wilson, 1995, p. 275).

[8] This is not the inference that Hirschberg predicts; rather she expects simply the exhaustive interpretation. However she also considers non-entailment scales such as fame of actors: "Did you get Paul Newman's autograph?" "I got Joanne Woodward's." The analysis we give simply combines the two notions.

[9] van Rooij and Schulz (2006) treat a similar type of example as a special case of exhaustive interpretation. Merin (1999), on the other hand, proposes treating standard scalar implicatures as special cases of conflict-of-interest examples like the ones discussed here.

of trustworthy communication. This section looks in more detail at further related cases of pragmatic inference beyond pure cooperation. The difference between the examples in this and those in the previous section is that the inferences discussed in this section go strictly beyond the classical cooperative (scalar) inferences discussed in the Gricean camp. The idea that messages are verifiable, i.e. that the speaker will speak truthfully, is maintained thoughout and only released in section 6.

4.1 'Unwilling to tell'-inferences

Let us first start by reviewing an example that has been argued to demonstrate a critical failure for neo-Griceans who rely too much on the Cooperative Principle. The example was originally discussed by Grice himself (Grice, 1989, example (3) on p. 32).

(5) A: Where does C live?
B: Somewhere in the South of France.

With this example, Grice showed how the ignorance implicature that B does not know where C lives can be derived straight-forwardly from the maxims of conversation and the cooperative principle: since B is less informative than required, but still cooperative, the only reason for not providing the information asked for is that she does not have that information.

To this example, relevance theorists have objected that there is another inference which can be drawn from B's utterance, if the context is slightly different (Sperber and Wilson, 1995; Carston, 1998). If it is common knowledge that B is knowledgeable about C's whereabouts, her answer implicates that she is *unwilling to tell*. Carston (1998) lists a number of nicely accessible further cases of such 'unwilling to tell'-inferences (her examples (63) to (65)):

(6) A: When will you be back?
B: When I'm ready.

(7) A: Which of your colleagues support the strike?
B: Some of them do.

(8) A: How many clients do you have?
B: A number.

The pattern of all these is fairly clear: the hearer would like to have certain information, but the speaker wants to keep the hearer uninformed (in related cases she might eschew the processing costs of *retrieving* the true answer). In any case, since the speaker could or should in principle know the answer to the hearer's question, the uninformative answer (perhaps with its grumpy tone) clearly communicates that the speaker is unwilling to give that information away.

This 'unwilling to tell'-inference is fairly natural, and, once we appreciate it fully, seems to be available in virtually *all* decontexutalized cases of attested scalar implicatures (Carston, 1998). However, these inferences are allegedly problematic for neo-Gricean accounts of implicatures, because, so Sperber, Wilson and Carston argue, they involve the hearer realizing that the speaker is purposefully or ostensibly *un*cooperative. This, the argument continues, cannot give rise to an implicature under the neo-Gricean conception, because implicatures have to be, by neo-Gricean definition, computed under the assumption of cooperation.[10] Relevance theory, on the other hand, arguably can account for this inference (Sperber and Wilson, 1995; Carston, 1998).

It is not entirely clear to us that the utterances giving rise to 'unwilling to tell'-inferences are so uncooperative that Gricean conceptions of pragmatic inference cannot account for them. After all, the speaker wants the hearer to draw this inference and we may assume that drawing this inference then is also in the interest of the hearer. The speaker is only partially uncooperative, intuitively speaking, not giving the hearer the information that he *really* wants.

A game-theoretic model of this inference is straight-forward, exactly because game theory can represent these fine-grained distinctions in agents' preferences. Here is a fairly simple-minded model with the sole intention of showing how game theory helps handle this case: suppose, for simplicity that there are only two possible places where C might be in the South of France, Cannes and Nice. Consequently we distinguish two states of the world $\texttt{Cannes}_t$ and $\texttt{Nice}_t$. The speaker does know where C lives and the hearer would like to know, i.e. the hearer would like to adopt a proper action which is optimal in either state of the world, so we have actions $\texttt{Cannes}_a$ and $\texttt{Nice}_a$. But now, since the speaker may or may not be interested in conveying the true state of the world, we should also distinguish states $\texttt{Cannes}_t^*$ and $\texttt{Nice}_t^*$ which are like the unstarred states, but where the speaker prefers to keep C's location a secret. To properly model the situation, therefore, we also need to make an interpretation action $?_a$ available for the hearer that is optimal exactly when the speaker does not want to tell where C lives — this action $?_a$ may represent the hearer adopting the belief that the sender does not want to tell him where C lives, or more concretely, deciding not to ask any more questions about C or similar. Notice, however, that the hearer wants first and foremost to know C's whereabouts; the action $?_a$ is strictly dispreferred to any concrete action $\texttt{Cannes}_a/\texttt{Nice}_a$ if the hearer knows the concrete state of affairs.

The signaling game this gives rise to is this:

[10] See for instance Carston (1998) who writes that the 'unwilling to tell'-inference "doesn't seem to be derivable at all using Grice's system. The problem is that it involves the hearer in recognising the absence of speaker cooperation and in his scheme, whatever maxims may be violated, the ultimate interpretation of an utterance must be such that the assumption of speaker compliance with the overarching Cooperative Principle (CP) is preserved. It can only be a case of what Grice calls "opting out", which does not give rise to implicatures at all."

	$\texttt{Cannes}_a$	$\texttt{Nice}_a$	$?_a$	$\texttt{Cannes}_m$	$\texttt{Nice}_m$	$\texttt{somewhere}_m$
$\texttt{Cannes}_t$	1,1	0,0	0,0	$\surd$	–	$\surd$
$\texttt{Nice}_t$	0,0	1,1	0,0	–	$\surd$	$\surd$
$\texttt{Cannes}^*_t$	0,1	0,0	1,.7	$\surd$	–	$\surd$
$\texttt{Nice}^*_t$	0,0	0,1	1,.7	–	$\surd$	$\surd$

Without going into formal details, it is certainly also intuitively appreciable that we can account for the 'unwilling to tell'-inference: the only reasonable equilibrium in this game (assuming, still, that the speaker has to be truthful) is where the speaker sends messages $\texttt{Cannes}_m/\texttt{Nice}_m$ in states $\texttt{Cannes}_t/\texttt{Nice}_t$ respectively and message $\texttt{somewhere}_m$ in both starred states; the hearer interprets messages $\texttt{Cannes}_m/\texttt{Nice}_m$ literally and draws the 'unwilling to tell'-inference in the form of action $?_a$ in response to message $\texttt{somewhere}_m$.[11]

4.2 Deceiving with the truth

In example (5) the 'unwilling to tell'-inference arose plausibly under the assumption that it was common knowledge between interlocutors that the speaker actually knew where C lived. But what if we are in a context where the speaker indeed knows where C lives, but does not want to reveal this information (which the receiver does not know), and where additionally the speaker believes that the hearer considers it possible that the speaker does *not* know where C lives? In that case the 'unwilling to tell'-inference does not suggest itself. Rather, we'd expect the hearer to draw the standard scalar inference as to the ignorance of the speaker. But this, of course, is a false inference; the whole case is an interestingly convoluted attempt of deception — notably a deception with a semantically true sentence whose normal pragmatic enrichment (assuming cooperativity) is false.

Here is another example of the same kind.[12] One fine morning, Robert comes to work and finds that Micha and Tikitu have eaten all of the cookies he had baked and brought to work the day before. So Robert is curious how many cookies Micha had (because Micha claims to be on a diet). Micha admits to having eaten *some* of them (since with crumbs all over his face, there is no point in denying it). Should Robert conclude that Micha did not eat all of the cookies, that Tikitu had some cookies as well? Not necessarily. In fact, in the situation given, despite his training in Gricean pragmatics, we could very well imagine that Robert does *not* draw the scalar inference "not all", because he knows that Micha would never have admitted having eaten all.

[11] There are also certain unreasonable equilibria in which the hearer responds to message $\texttt{somewhere}_m$ with either $\texttt{Cannes}_a$ or $\texttt{Nice}_a$. That standard equilibrium notions are too weak to respect (our intuitions) about semantic meaning is a topic that we will come back to in section 6 where we sketch an alternative solution concept that does select only the intuitively reasonable behavioral pattern in this game.

[12] The present example is a variation on an example discussed by Green (1995) and Carston (1998). Yet, the variation we are looking at —attempted deception with semantically true messages by exploiting cooperative pragmatic inferences— is, to the best of our knowledge, new.

The example takes the following (simplified) form: Micha has a choice between messages $\texttt{some}_m$ and $\texttt{all}_m$, but strictly wants Robert to believe that he only ate some of the cookies. Robert on the other hand, wants to know the true state of affairs.[13]

	$\exists\neg\forall_a$	$\forall_a$	$\texttt{some}_m$	$\texttt{all}_m$
$\exists\neg\forall_t$	1,1	0,0	$\surd$	–
$\forall_t$	1,0	0,1	$\surd$	$\surd$

And indeed, if these payoffs are commonly known —and if we do not assume any external preferences for/against non-misleading signals— the scalar implicature should *not* arise in this case, simply because sending the stronger alternative (and having it believed), though available and *hearer*-relevant, would go strictly against the sender's preferences. So, from the interpreter's perspective, the meaning of a scalar item should not be strengthened if the speaker cannot be expected to have used an alternative form (given her own preferences).

So far, this verdict is very similar to example (5) and the 'unwilling to tell'-inference, but still not quite the same. We admit that it is possible to assume that Micha intended to convey an 'unwilling to tell'-inference. But it is also possible, and indeed much more plausible, to imagine that Micha hoped or believed that Robert would strengthen "some" in the usual way to "some but not all". Then he can hope, if not expect, that his answer would *mislead* Robert (despite being true!) into concluding something false by a pragmatic inference (one that is usually called for but not, as it happens, in the concrete situation at hand).

Cases where agents try to mislead with the truth by exploiting an expected pragmatic inference are already interesting enough on their own.[14] But we can

[13] The structure is very similar to the "seduction game" we used to introduce persuasion games. The difference is that in the seduction game the low-quality speaker is forced to reveal her quality; both types want to use the same message, but one is prevented from doing so. This is why in the seduction game, unlike in this example, naivety unravelling gives rise to a fully revealing equilibrium.

[14] The question whether a witness is guilty of perjury if she (deliberately or not) uses a pragmatically misleading, yet semantically true statement is an issue of foremost legal relevance. Consider, for instance, the *Bronston case*, a seminal court case in matters of perjury and literal truth in U.S. legal history (see Solan and Tiersma, 2005, chapter 11; this reference was brought to our attention by Chris Potts and we are very grateful for that). Samuel Bronston, president of a movie company, was tried for perjury based on the following statements he had made under oath during the bankruptcy hearing for his company (from Solan and Tiersma, 2005, p. 213):

(1) Questioner: Do you have any bank accounts in Swiss banks, Mr. Bronston?
Bronston: No, sir.
Questioner: Have you ever?
Bronston: The company had an account there for about six months, in Zurich.

In fact, as was later revealed, Bronston *did* have a personal Swiss bank account with a substantial sum of money during a period of several years relevant to the bankruptcy hearing. He was found guilty of perjury, because his statement, though literally true, falsely implied that he did not have any Swiss bank accounts. However, the U.S. Supreme Court reversed also take the whole idea one step further and look at the interpreter's side: not only can speakers try to mislead (by exploiting whatever mechanism or belief configuration), hearers can in principle see through attempts of deception. But if they do, this too should be considered a pragmatic inference (in a loose sense of the term, if you wish): it is an inference based in part on the semantic meaning of the utterance and establishes reasons and motives why this utterance has been produced.[15]

Of course this inference, much like the scalar implicature in connection with example (1), is not one that the speaker *wants* the hearer to draw. In the scenario, the speaker intends the hearer to draw the scalar implicature as a means of deceiving him; we would like to say that she probably *does* want the intention of communicating the scalar inference to be recognized, but she does *not* want her deceptive intent to be recognized (as such). However, if the hearer takes the speaker's own interests rightfully into account, he may see through this deceptive intention and *not* draw the scalar inference. The crucial realization here is that this is higher-order reasoning *on top of ostensive as-if-cooperative communication*, because in order to see through the deception first the scalar inference has to be computed.

To sum up here, we think that these examples of 'pragmatic deception' and their recognition are worth the pragmaticist's attention for at least two reasons. First of all, it should be apparent that the Gricean maxims do little to lead the hearer to the realization that he is being deceived. In the next section we argue that Relevance Theory —a long-term competitor of neo-Gricean pragmatics which is slightly less dependent on cooperativity— also has its problems explaining the recognition of attempted deception. Secondly, the discussion above showed that there are at least three levels of sophistication in hearer reaction to attempts of deception by pragmatic enrichment: blindly naive speakers would perhaps take a message "some" entirely at (semantic) face value; it arguably takes some pragmatic skill to compute scalar implicature; but only even more sophisticated hearers understand that the speaker is trying to outsmart them and by this realization they, in turn, can outsmart the speaker. We will explore this kind of reasoning —step-wise strategic reasoning that takes into account

the verdict, arguing that the negative implication of Bronston's answer should have been recognized as an implication, not as literally stated, and that it would have been the questioning lawyer's responsibility to probe for clarity.

The Bronston case is very relevant to the matters discussed in this paper. The inference under discussion is very much parallel to example (4) ("Do you speak Portuguese?" "My wife does.") discussed in section 4.1, except that there the speaker would have wanted to give an affirmative answer to the direct question that was posed, whereas Bronston obviously did not. We can see clear practical reasons for the Supreme Court's decision, but we are also happy to point out a parallel to the very point we are trying to make here, namely that it takes a *higher* level of sophistication to see through an attempted 'pragmatic deception' than to compute the pragmatic inference the attempted deception exploits, but that to be on the guard for unwarranted presumptions of linguistic cooperation is, in extreme cases, a rational imperative.

[15] To see through a deception is, if you want to look at it like this, much like establishing what *speech act* has been performed: this was not an informative assertion, but an attempt at deceit or even a lie.

the speaker's self-interested strategy— further in Section 6.

5 Deceit, relevance and forward induction

The Gricean maxims of conversation give, on the face of it, no direction how the hearer is to arrive at the fine-grained interpretation that he is being deceived by an uncooperative speaker. This holds true not only for most, if not all neo-Gricean elaborations on the Gricean maxims, but also for *post-Gricean*[16] Relevance Theory, although the latter, as we will see in a moment, incorporates the speaker's interests to some extent in pragmatic interpretation. This section is dedicated to the argument that Relevance Theory does not incorporate *enough* of the speaker's interests in its interpretation principles. We compare this verdict with what we consider a close-enough game-theoretic analogue of a relevance-based interpretation principle: the Best Rationalization Principle, which will serve as the backbone of the model to be sketched in section 6.

5.1 Inference based on the presumption of optimal relevance

Relevance theory (Sperber and Wilson, 1995, 2004) holds that the central ingredient in the interpretation of ostensive stimuli, which have an obvious communicative intent, is the assumption that the stimulus to be interpreted is *optimally relevant.* For linguistic utterances this yields the following formulation of the

Communicative Principle of Relevance: Every utterance communicates a presumption of its own optimal relevance (Sperber and Wilson, 1995, p.260).

Optimal relevance, in turn, is defined as follows:[17]

Optimal Relevance: An utterance is optimally relevant to an addressee iff:

(i) it is relevant enough to be worth the addressee's processing effort;

(ii) it is the most relevant one compatible with the speaker's abilities and preferences (Sperber and Wilson, 1995, p.270).

[16] We adopt Carston's (1998) useful terminology here which contrasts the neo-Gricean pragmatics of Atlas and Levinson (1981); Levinson (1983); Horn (1984) and others with *post-Gricean* pragmatics. While both strands of research try to explain pragmatic interpretation as an inference, the neo-Griceans seek to explain pragmatic inferences as based on explicit formulations of Grice's maxims of conversation and, in particular, the Cooperative Principle. Post-Gricean theories do not rely on Grice's maxims and the Cooperative Principle. Both relevance theory and the kind of game-theoretic approaches to pragmatic inference we are advocating are post-Gricean in this sense.

[17] This is crucially the revised version of the definitions offered in the postface of Sperber and Wilson (1995).

For our present purposes it is crucial to note the explicit reference to the speaker's abilities and preferences in the second clause of the definition of optimal relevance. Nevertheless, it seems that relevance theory still has a clear focus on *hearer* relevance and as such obscures fine-grained inferential distinctions based on speakers who are willing to be forthcoming to different degrees. To be precise, we do not see how the presumption of optimal relevance, formulated as above, can guide the hearer in his realization that he is (possibly) being deceived, as discussed above in section 4.2, by exploitation of a pragmatic inference. The problem is that the speaker's preferences only feature as a blocking mechanism, requiring that a potentially hearer-relevant interpretation must also conform to the speaker's preferences.[18] This does not suffice to explain a speaker's entirely self-interested, deceptive behavior. In other words, relevance theory does not carry far enough away from Grice's Cooperative Principle: it is unclear what interpreters may infer based on optimal relevance when speakers clearly have an incentive to mislead and cheat.

On the other hand, game theory, which after all is the analysis of conflict between self-interested agents, is much more at home in these wild waters. In order to work out the difference in perspective, and to build towards the model presented in the next section, we would like to work out a rough game-theoretic analogue of the relevance theoretic presumption of optimal relevance and briefly compare the two.

5.2 Forward induction and the best rationalization principle

We believe that the game-theoretic notion of *forward induction* is closely related to the notion of *relevance* in natural language interpretation.[19] As a motivating example of forward induction reasoning, consider the famous Hawk-Dove game. Here, the speaker is the row player and the hearer is the column player; both of them have a choice between playing hawk (h_S and h_H respectively) and dove (d_S and d_H; as usual, the row player's payoff is given first).

The original Hawk-Dove game

	h_H	d_H
h_S	$-2, -2$	$2, 0$
d_S	$0, 2$	$1, 1$

Intuitively, these payoffs represent that it is most desirable for a player to play hawk if the other one plays dove (the hawk player gets the full cake, the dove player gets nothing); but if two hawk players meet, they destroy each other and

[18] We assume here that the locution "the most relevant" in clause (ii) of the definition of optimal relevance refers to hearer relevance, as we have called it. There is no discussion of this potential ambiguity in Sperber and Wilson (1995), but an interpretation as speaker relevance makes little sense given the qualification to speaker's abilities and preferences.

[19] Seminal work on forward induction is by Kohlberg and Mertens (1986) and van Damme (1989). For more on the relation between forward induction and natural language interpretation see Sally (2003).

that is far worse than when two dove players meet and peacefully share the resource at stake.

This game has two (asymmetric) Nash equilibra in pure strategies, $\langle h_S, d_H \rangle$ and $\langle d_S, h_H \rangle$, and it is clear that each individual prefers one equilibrium above the other. Without any commonly shared expectations what the other player will do, it is simply not clear for each player what is the best action to choose. But the game also has a Nash equilibrium in mixed strategies, where both players play Hawk with probability $\frac{1}{3}$. The expected payoff for each player if the mixed Nash equilibrium is played is $\frac{2}{3}$.

Consider now the following variant of this game, where the speaker S has the opportunity to inflict some damage *on herself* prior to playing the familiar (static) Hawk-Dove game against the hearer H (Ben-Porath and Dekel, 1992; Shimoji, 2002). Assuming that this damage equals one util for S this gives rise to the following dynamic game, where S begins by playing a_1 or a_2:

a_1, do nothing

	h_H	d_H
h_S	-2,-2	2,0
d_S	0,2	1,1

a_2, selfdamage

	h_H	d_H
h_S	-3,-2	1,0
d_S	-1,2	0,1

Why should a rational agent choose to hurt herself? Indeed, *backward induction* predicts that she should not. Backward induction is an iterative procedure that determines each moving player's optimal choices in each subgame of a dynamic game, starting from the last choice points where players move and then propagating optimal choices backwards —hence the name— to all earlier choice points. Here's what backward induction does in the above dynamic Hawk-Dove game. This game has two (strategic) subgames which are strategically equivalent: they both have the same Nash equilibria (pure $\langle h_S, d_H \rangle$ and $\langle d_S, h_H \rangle$ and mixed with $P(h_S) = P(h_H) = \frac{1}{3}$). The only difference between the two subgames is that S's expected payoff from any equilibrium, pure or mixed, is exactly one util less in the game after a_2 than in the one after a_1. But that means that if the hearer makes his choice in both subgames independent of whether the speaker chose to hurt herself or not, it would indeed be irrational for the sender to do so. Backward induction predicts exactly that, because backward induction —the name is somewhat unfortunate when we look at things in this way— only looks *forward* into the future moves of the dynamic game and does not take into account the previous game history that led to a particular subgame.

Still, there is ample reason why it may be rational for S to play a_2 after all. S might believe that H would choose to play d_H if he observes her hurting herself, but would otherwise play h_H with some positive probability. In that case, it is absolutely rational from S's point of view to inflict damage on herself, because if H indeed plays dove after a_2, S actually gains by self-sacrifice after all, because she can play hawk and expect a payoff of 1, whereas after playing a_1 her expected payoff is strictly smaller than 1.

This is where forward induction reasoning enters. In proper subgames of a dynamic game, unlike backward induction, forward induction recommends to look *back* —again the unfortunate naming— at the history of the play that led to the given subgame and to try to rationalize the past behavior that led there (if that is possible). There are many different versions of this idea in the economics literature and it is fair to mention that there is no consensus as to what formal notion satisfactorily captures this intuitive reasoning in its entirety. The most accessible approach to forward induction reasoning is, perhaps, via the epistemic *Best Rationalization Principle* (Battigalli, 1996, p.180):[20]

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

For our purposes here, we do not want or need to go into any more formal detail; the informal formulation of the Best Rationality Principle suffices to grasp sufficiently the idea of forward induction. The idea is that all past behavior is to be rationalized as much as possible: we must believe that others are maximally rational given the choices that they have made, even when they fail to choose what seems to us the most rational option.

5.3 Inference, relevance and forward Induction

It should be clear how the Best Rationalization Principle gives rise to the intuitive verdict above that H should believe that S believes that H plays d_H after she has hurt herself with a_2, since in the particular example, this belief of S is the *only* belief ascription which has self-damage come out rational. What is true for arbitrary actions is true for linguistic actions as well. Indeed, we might look at actions a_1 and a_2 either as differently costly messages (whatever would motivate the costs) or, what is perhaps even more suggestive for natural language interpretation, we might regard action a_1 as a costless "not doing or saying anything", whereas the cost of action a_2 is merely the production cost of whatever ostensive stimulus might attract the hearer's attention.[21]

Thus conceived, forward induction reasoning can serve as a pivotal element also in ostensive communication: if S knows that H will rationalize apparently self-destructive behavior, she can thereby communicate very safely her intent of playing hawk, because this, to repeat, is the only belief that has a_2 come out rational. We would like to suggest that this idea of rationalization *ex post* in dynamic games can be seen as a game-theoretic explication of relevance-based

[20] This principle has been implemented in epistemic models of games as *robust or strong belief in rationality* (Stalnaker, 1998; Battigalli and Siniscalchi, 2002). This latter notion is central for epistemic characterizations of both rationalizability (Bernheim, 1984; Pearce, 1984), iterated weak dominance and also of the *intuitive criterion* of Cho and Kreps (1987).

[21] However, it is absolutely essential to understand that the idea of forward induction, and with it the application of the Best Rationalization Principle, does not depend at all on messages having differing costs attached to them; this was just a feature of the example with which to introduce the intuitive notion most perspicuously.

natural language interpretation. To establish the meaning of an apparent *non-sequitur*, for instance as in the classical example (9) of Grice (Grice, 1989, p. 32), the hearer needs to rationalize the speaker's linguistic behavior, especially where it deviates from expectation; the question to be asked and answered is: "which beliefs of the speaker justify best that she said *that*?"

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

Despite the (at least crude and superficial) parallels, there are crucial differences between the presumption of optimal relevance entertained in relevance theory and the Best Rationalization Principle, which favor the latter (at least) as an explanation of linguistic behavior in cases of conflicting interests. Consider again the cookies example from section 4.2, where Micha sought to mislead Robert with the statement that he had eaten *some* of the cookies. The Best Rationalization Principle gives the hearer a guide to infer exactly this. The hearer asks "which strategy profile best rationalizes the sent message $\texttt{some}_m$?", and finds *one* such best rationalization in the possibility that the speaker expects him to draw the scalar implicature, and then abuses this inference for her personal gain sending $\texttt{some}_m$ in state $\texttt{all}_t$. The Best Rationalization Principle handles this sort of pragmatic inference beyond cooperative ostensive communication with ease. Optimal relevance does not easily deal with this case, as we have argued above.

To conclude this section, we suggest forward induction reasoning as a game-theoretic analogue of relevance-based natural language interpretation. In interpreting utterances, hearers have to rationalize the observable action, including the fact that a possibly costly or surprising signal was produced. This rationalization places the sender's payoffs center stage and thus makes entirely clear that from a game-theoretic perspective the notion of relevance at stake in natural language interpretation is speaker relevance. Although in most cases the interests of the speaker might align with the interests of the hearer, this is due to some further and often external motivations feeding into the speaker's own preferences as argued in Section 2.

In the following section we investigate cases of pragmatic inference when we drop the assumption that messages are verifiable and speakers speak truthfully. The model we suggest to cope with this situation shows how the deeper nestings of strategic reasoning required to see through certain cases of deception can come about based on the Best Rationalization Principle.

6 Pragmatic inference beyond truthfulness

6.1 Cheap talk and credibility

In sections 3 and 4 we investigated pragmatic reasoning in the light of an apparent conflict of interest between speaker and hearer. Still, we assumed that speakers *have to* speak truthfully, and we motivated this crude assumption with the idea that messages are verifiable: that the hearer (or some external judge) has a way of checking for truth and punishing the speaker for untruthfulness. However, in real life, situations are readily conceivable where the personal preferences of speakers outweigh any inclination towards cooperation, and in particular towards truth. Contrary to the assumption of verifiability, we can easily imagine cases where hearers can (or would) never check on the truth of the information. Just think of the myriads of small, unimportant face-saving lies told everyday: saying “My bus was delayed” as an apology when in fact you were lazy and just got up too late.

The question arises, when is semantic meaning credible? This is arguably an issue for linguistic pragmatics too (see below), and we could credit a game-theoretical approach to the field for making these questions heard and easily intelligible. (Whether you prefer them answered in game-theoretic models is another issue.) In approaching this issue, the limit case of entirely unverifiable speech, devoid of all psychological or social incentives for cooperation and truthfulness, is particularly interesting from a modelling perspective and has received due attention. This constellation is known as *cheap talk* in the economics literature (Farrell and Rabin, 1996). Talk is cheap, in a signaling game, if every possible message $m \in M$ may be used in every state, all incurring the same costs, so that the sender’s payoffs only depend on the state t and the receiver’s response a.[22] We will use the idea of cheap talk to motivate the investigation of MESSAGE CREDIBILITY in the following. It should be noted though that the model we sketch subsequently to address this issue does not *require* cheap talk, but is flexible enough to handle verifiable and costly messages equally well.

To see that we do have rigid intuitions about message credibility, consider a simple arranged situation. Ann and Bob are playing the following game. A judge flips a fair coin and only Ann observes the outcome of the coin flip, but Bob doesn’t. Bob has to guess the outcome of the coin flip and wins iff Ann loses iff Bob guesses correctly. But suppose that before Bob makes his guess, Ann 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.[23] This is, in effect a ‘matching pennies’-style, zero-sum signaling game with cheap talk of the following form:

	heads$_a$	tails$_a$	heads$_m$	tails$_m$
heads$_t$	0,1	1,0	$\surd$	–
tails$_t$	1,0	0,1	–	$\surd$

How should Ann’s announcement affect Bob’s decision? It seems it shouldn’t at all. Bob knows that Ann 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 Ann says in his guess.

[22] Notice that cheap talk does not, strictly speaking, mean that signals incur *no* costs. Signaling may be heavily expensive under this definition as long as all signals in M incur the same cost.

[23] We could have Ann 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.

But now, consider a slightly adapted version of the above game. Suppose that while Bob is out of the room, either the coin is flipped or a judge tells Ann that it's "share time". If it's share time, and Bob guesses correctly that it is, both Ann and Bob win. But if Bob guesses on a coin flip outcome when it's share time (or vice versa), both Ann and Bob lose. And, moreover, Ann can now additionally announce that it's share time whenever she wants without constraints as to the truth. The resulting signaling game looks like this:

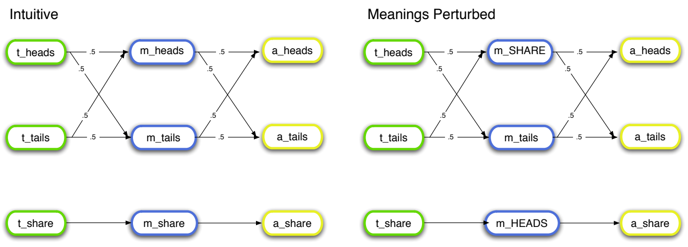
	$\texttt{heads}_a$	$\texttt{tails}_a$	$\texttt{share}_a$	$\texttt{heads}_m$	$\texttt{tails}_m$	$\texttt{share}_m$
$\texttt{heads}_t$	0,1	1,0	0,0	$\surd$	–	–
$\texttt{tails}_t$	1,0	0,1	0,0	–	$\surd$	–
$\texttt{share}_t$	0,0	0,0	1,1	–	–	$\surd$

Just as in the simpler signaling game above, the messages $\texttt{heads}_m$ and $\texttt{tails}_m$ are intuitively *not* to be believed. But interestingly, the message $\texttt{share}_m$ that it is share time, is (very) credible. (Put differently: the resulting game can be separated into a zero-sum portion and a coordination portion; messages that intuitively relate only to the zero-sum portion should be disregarded; messages regarding the coordination portion should be taken seriously.)

The simplicity of the example and its intuitive plausibility might obscure the appreciation of a conceptually very important point: there are very robust and natural intuitions about the reliability of given meaningful signals in a given set of (admittedly stylized) strategic situations. As with previously discussed cases of attempted deception, we suggest that these intuitive judgements are rightfully labelled pragmatic inferences (again, in a loose sense of the term if you insist), as they are inferences based on the semantic meaning and the context of utterance, in particular and crucially the motives of the speaker.

In fact, game theorists have addressed this issue, saying that neither message $\texttt{heads}_m$ or $\texttt{tails}_m$ is *credible*, but that $\texttt{share}_m$ is: intuitively, Bob should believe only credible messages (Rabin, 1990; Farrell, 1993; Farrell and Rabin, 1996). 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 the semantic meaning of the available messages. Intuitively speaking, Ann's message $\texttt{share}_m$ is credible and will be believed, because *its* semantic meaning is indicative of a situation where payoffs are sufficiently (in fact, totally) aligned. But in cheap talk signaling games where semantic meaning is not binding, this is not readily explained without further effort. After all, if coordination is all that matters, Ann and Bob could also use the signal $\texttt{heads}_m$ to faithfully communicate that it's share time: to wit, not only is the strategy profile in figure 1a an equilibrium of the game, without binding semantic meaning the strategy profile in figure 1b where Ann randomly chooses either message $\texttt{share}_m$ or $\texttt{tails}_m$ with equal probability in states $\texttt{heads}_t$ and $\texttt{tails}_t$ and uses $\texttt{heads}_m$ if it is share time is also an equilibrium.[24]

[24] The strategy profiles in figures 1a and 1b are not the only equilibria in the cheap talk game, of course. Most importantly, the sender does not have to send messages in states $\texttt{heads}_t$ and $\texttt{tails}_t$ with probability exactly .5, as long as she plays the same strategy in both states.



(a) Intuitive Equilibrium

(b) Crooked Equilibrium

Figure 1: Equilibria in the Extended Matching Pennies Game

It might seem that we could solve this problem by introducing (uniform) costs for untrue messages, however this is not the case. Taking the influence of semantic meaning into account requires a delicate balance: the semantic meaning must not be binding (or there is no possibility of lying for strategic gain) but it must not be completely irrelevant either. Costly signalling cannot strike this balance: if costs are too small then permuted equilibria survive (those in which the state $\texttt{share}_t$ is indicated by one of the other messages); if costs are large enough to eliminate these 'perverse' equilibria then they rule out strategic lying as well.[25]

6.2 Iterated best response reasoning with focal points

How do we then implement both a tendency to stick to semantic meaning where this is compatible with the strategic situation *and* some version or other of the Best Rationalization Principle? The model that we would like to sketch in this section addresses exactly this concern. We will here only sketch the common core of a class of models that, with differing details and ambitions, have been suggested for pragmatic interpretation and whose main idea is that of (iterating) best responses with semantic meaning as a focal starting point (Benz, 2006; Stalnaker, 2006; Jäger, 2007; Benz and van Rooij, 2007; Franke, 2008; Jäger, 2008).

6.2.1 Motivating semantic meaning as focal starting points

So, what is iterated best response reasoning with focal points? Consider the following simple 'hide-and-seek' game. There are four labelled and linearly arranged doors, as shown here:

[25] The problem has been addressed in the equilibrium refinement literature; for an overview, see Farrell and Rabin (1996). Unfortunately there are problems, technical and conceptual, that prevent a direct application of existing notions of message credibility from economics to linguistic pragmatics (Franke, 2008). The most significant is that many economic approaches assume total effability: that every possible meaning be expressible by at least some message in the game. This is decidedly *not* what we want for linguistic pragmatics: to wit, scalar implicatures arise only given at least partial ineffability.

A B A A

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

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

When looking at the game in this abstract form, there is nothing *in the payoffs* that should prejudice any of the four doors over any other for *Hider*, nor for *Seeker*. However, the different labeling of doors and their linear arrangement does seem to make a difference to human reasoners. When Rubinstein et al. (1996) put this condition to the test, they found that the following percentage of subjects chose the various doors:

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

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

Experiments following this paradigm have been multiply replicated. Surveying these, Crawford and Iriberri (2007) argue that Rubinstein et al.'s empirical results in this and similar 'hide-and-seek' games on *non-neutral landscapes* (their term for the psychological saliency effects exemplified above) can best be explained by assuming that there are focal points: the door labeled B and the left- and right-most doors attract attention in the sense that this is the first spot where a prize would be hid and sought. But then, starting from this initial focal prejudice, human reasoners may show different levels of strategic sophistication and reason themselves away from the focal point, so to speak, in the following way. A *Hider* might, for instance, reason: "Suppose I hid the prize behind the B-labeled door (which is a very attractive choice for reasons I don't quite understand), then maybe this is also an attractive choice for *Seeker*, so I should *not* hide it there. I should maybe go for the left-most door (also pretty appealing somehow, strangely enough), but then *Seeker* might anticipate this so I should rather ..."[26]

Such focal point reasoning is fairly intuitive also for models of natural language interpretation. Given a semantically meaningful message, the hearer would like to rationalize why the speaker said what he said. The semantic meaning of the perceived message is a focal point, we suggest, much like the door labeled B: even though strategically semantic meaning is not binding, it is fairly intuitive to start pondering what might have been meant by assessing the semantic meaning. So, as a starting point of his deliberation the hearer asks himself, what he would do if the message was indeed true. But then he might realize that the sender could anticipate this response. In that case, the hearer is best advised to take the strategic incentives of the sender into consideration. The resulting hypothetical reasoning on both the sender and the hearer side can be modelled as a sequence of iterated best responses.

Fully fleshed-out models of such reasoning in the context of natural language interpretation have been suggested recently by Franke (2008) and Jäger (2008). We will give a largely informal sketch of these models in the following.

6.2.2 Iterated best response reasoning as pragmatic inference

IBR models intend to capture actual human reasoning behavior with its natural resource-bounded limitations. To achieve this, the main idea in IBR models of human reasoning is to distinguish different *strategic types* of players (Stahl and Wilson, 1995; Crawford, 2003; Camerer et al., 2004). A strategic type captures the level of strategic sophistication of a player and corresponds to the number of steps that the agent will compute in a sequence of iterated best responses. More concretely, the starting point of the IBR sequence is given by unstrategic, even possibly irrational, level-0 players. A level-$(k+1)$ player then plays a best response to the behavior of a level-k player. (A *best response* is a rationally best reaction to a given belief about the behavior of all other players.)

As for signaling games, a level-0 player represents a naive hearer (who simply takes messages as true) or speaker (who chooses a true message at random). These two types represent the assumption that semantic meaning is focal. We can generate inductively a hierarchy of strategic types: a level-$(k+1)$ hearer plays a best response to a level-(k) speaker, and so on.[27]

[26] It perhaps deserves mention for the sake of clarity: in IBR models focal points play a role different from, both technically as well as conceptually, and not to be confused with Schelling's notion of focal points as criteria for selecting among multiple equilibria (Schelling, 1960). Of course, the problem of selecting the intuitive equilibrium in the above extended matching pennies game could well be thought of as an application of Schelling's idea of focal equilibria, but in general the approach taken by IBR models with focal points is different, as evidenced by the 'hide-and-seek' game with differently labeled doors: there is no set of equilibria to select the focal ones from, but still focal elements play a role in the explanation of actual human reasoning in this case.

[27] We gloss here over a number of subtleties that are explored in the papers referenced above. In particular, a high-level hearer need not (as we assume here for simplicity) believe he faces a speaker of the next level below, but might have more complex beliefs about likelihoods of speakers of all lower levels. In addition, the hierarchies founded on a naive speaker and hearer differ but are interrelated; in our discussion we will assume we have access to both.

We can think of climbing the iterated best response hierarchy as a process of successive approximation to idealised rational behaviour, as predicted by the best rationalisation principle. The strategies that are consistent with the entire infinite hierarchy provide a solution concept for idealised agents, but the finite levels represent more realistic resource-bounded, agents. We can define credibility of messages in the abstract, for instance, by looking at the entire hierarchy: in the extended matching pennies game we introduced in Section 6.1, no speaker at any level will have an incentive to use the message "It's share time" untruthfully (matching our intuition that the message is credible), while "Heads!" and "Tails!" are used to lie to a naive hearer at the first level.

What the IBR model achieves is the right balance of the influence of semantic meaning: influential (via focal meaning) but not overwhelming (the balance that we argued in Section 6.1 could not be achieved by message costs alone). It is easily verified that the strategy profile in figure 1a is the IBR model's sole prediction of perfectly rational behavior in the extended matching pennies game.[28] Moreover, the same model makes good predictions for a variety of more standard examples where preferences are (by and large) aligned, such as various types of implicatures (see Jäger, 2008). At the same time, the notion of a bounded strategic type lets us cope with cases of deception and the recognition thereof; exactly the kinds of cases that we argued are problematic for Gricean, neo-Gricean and relevance theoretic pragmatics.

6.2.3 Deception and outsmarting

The IBR model is founded on the assumption that a player of high strategic type always believes that his opponents are of lower type (that is, that they are less sophisticated strategic reasoners). This self-confident assumption of strategic superiority may seem unreasonable, even irrational, but experimental studies unfortunately show that it is quite realistic (Ho et al., 1998; Camerer et al., 2004). This trace of overconfidence in the model also proves helpful to cope with cases of attempted deception and the recognition thereof.

Overconfidence is needed in any explication of attempted deception —be it a failure or success—, because some imperfection —be that in terms of false or partial beliefs, or other severe cognitive limitation— must exist for deception to have any hope of success: if speaker and hearer are perfectly informed about each other's possible moves and payoffs, and if they are perfectly rational reasoners with common belief in rationality, any attempt of deception will be anticipated and the anticipation thereof will be anticipated as well and so on ad infinitum (see Ettinger and Jehiel, 2008). The upshot of this argument is that wherever we do see attempted deceit in real life we are sure to find at least a belief of the deceiver (whether justified or not) that the agent to be deceived has some sort of limited reasoning power that makes the deception at least conceivably

[28] Similarly, the IBR model yields the intuitive prediction also in the Cannes/Nice game of section 4.1 and accounts thus for the 'unwilling to tell'-inference in connection with example (5). It also seems to replicate the results of section 3 on persuasion games with verifiable messages (although this requires again an ϵ-probability of naive interpretation error).

successful. In the majority of cases, perhaps, this limitation is likely just a run-of-the-mill ignorance that the speaker may have deviant preferences (i.e., preferences not aligned with the hearer's). But it may also be that we are misled, when we do in fact know the preferences of our conversation partners, but we fail to take the strategic ramifications of the entire situation sufficiently into account.[29]

This latter situation is accounted for by the IBR model of pragmatic reasoning. Take, for instance, the example we considered in Section 4.2 where Micha admitted to having eaten some of the cookies. The IBR model can account for the deeper nestings of strategic sophistication in this example. A naive, level-0 hearer would simply believe the message according to its semantic meaning and assume that the true state of the world is either $\texttt{some}_t$ or $\texttt{all}_t$. A level-1 hearer, on the other hand, at least takes the structure of the set of available messages and their semantic meanings into account (he assumes a naive speaker, who is indifferent between true messages). He would respond to message $\texttt{some}_m$ with action $\texttt{some}_a$ only, and not with $\texttt{all}_a$.[30] But then a level-2 sender will actually send message $\texttt{some}_m$ expecting (according to his belief that his audience is a level-1 hearer) that the scalar implicature will be computed. This models an attempted deceit (with a semantically true message): based on his belief that this message will induce the wrong interpretation, Micha sends it to try to improve his own payoff. However, the level-3 hearer, who in turn outsmarts the level-2 sender, believes that Micha is doing exactly this: sending message $\texttt{some}_m$ in the expectation that this will trigger the interpretation $\texttt{all}_a$. This deception is recognized on this higher level of reasoning and our level-3 hearer Robert will in fact respond to message $\texttt{some}_m$ indifferently with both interpretation actions.

7 Conclusions

We have come a long way, with this kind of strategic second-guessing, from Gricean cooperation. Nonetheless, the main suggestion that we wish to make with this paper is that this kind of reasoning deserves to be called pragmatics, and to be studied using the same tools that we use to investigate scalar implicature, the garage around the corner, and all the familiar examples. These are all inferences about how utterances can be interpreted beyond their semantic meaning, and they are all based on strategic reasoning about speaker motivations and choices. In particular, we have suggested that more attention needs to be paid to 'speaker relevance', to the preferences of the speaker as distinct from (and possibly in conflict with) those of the hearer.

In Section 2 we gave a suggestion as to why we should expect speaker preferences to be largely, but not universally, aligned with those of hearers. Sections 3

[29] Think, for instance, of the Bronston courtroom case mentioned in footnote 14. By all means, the attorney must have known that Bronston had no incentive whatsoever to admit that he had a Swiss bank account.

[30] To be perfectly accurate, this only holds as long as he has no strong prior beliefs that $\texttt{some}_t$ is more likely. We omit the details.

and 4 considered the (intuitively common) case where verifiability of messages constrains speakers to using only semantically true utterances (an analogue of the Gricean quality maxim), but where preferences still diverge in that speakers would prefer not to reveal all their information. We saw that under such constrained circumstances some inferences can still be drawn: a scalar-like inference that the speaker's choice is the true message which reflects best on herself (which in certain cases coincides with the semantically strongest, but need not always do so), and an inference that the speaker is simply unwilling to give more information.

The 'unwilling to tell' inference raises problems for neo-Gricean accounts, and the recognition of deceitful intent appeared as a problem also for relevance theory (as we've argued in Section 5). A solution can be found in game theory, which explicitly represents a speaker's preferences as the foundation of her strategic choice of action. However standard game-theoretic analyses either fail to adequately represent the influence of semantic meaning, or introduce it as a hard constraint which unduly limits the possibilities for deceitful language use. In the last section we gave a brief overview of the iterated best response model, which incorporates semantic meaning as a focal starting point but allows strategic reasoning about pragmatic implicature (the standard cooperative case), more extended inferences to do with deceitfulness, and even outright lies.

The focus on speaker relevance raises a number of interesting questions. What incentives induce speakers to align their preferences with those of their interlocutors? Under what circumstances are messages credible, and when are we justified in going beyond their semantic meaning? The main shift in perspective it produces, though, is an expansion in what is properly considered pragmatic inference. Through considering the speaker as an independant agent pursuing her own particular goals, we bring a wider range of interesting linguistic behaviours into the same general framework that we are accustomed to using for cooperative pragmatic reasoning.

References

Ariel, M. (2004). Most. *Language*, 80(4):658–706.

Atlas, J. D. and Levinson, S. (1981). It-clefts, informativeness, and logical-form. In Cole, P., editor, *Radical Pragmatics*, pages 1–61. Academic Press.

Battigalli, P. (1996). Strategic rationality orderings and the best rationalization principle. *Games and Economic Behavior*, 13:178–200.

Battigalli, P. and Dufwenberg, M. (2007). Dynamic psychological games. to appear in Journal of Economic Theory.

Battigalli, P. and Siniscalchi, M. (2002). Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106:356–391.

Ben-Porath, E. and Dekel, E. (1992). Signaling future actions and the potential for sacrifice. *Journal of Economic Theory*, 57:36–51.

Benz, A. (2006). Utility and relevance of answers. In Benz, A., Jäger, G., and van Rooij, R., editors, *Game Theory and Pragmatics*, pages 195–219. Palgrave.

Benz, A., Jäger, G., and van Rooij, R., editors (2006). *Game Theory and Pragmatics*. Palgrave McMillan.

Benz, A. and van Rooij, R. (2007). Optimal assertions and what they implicate. *Topoi*, 26:63–78.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, 52(4):1007–1028.

Bowles, S. and Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, 65(1):17–28.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898.

Carston, R. (1998). Informativeness, relevance and scalar implicature. In Carston, R. and Uchida, S., editors, *Relevance Theory: Applications and Implications*, pages 179–236. John Benjamins, Amsterdam.

Cheney, D. L. and Seyfarth, R. M. (1990). *How monkeys see the world*. University of Chicago Press.

Cho, I.-K. and Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221.

Crawford, V. P. (2003). Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions. *American Economic Review*, 93(1):133–149.

Crawford, V. P. and Iriberri, N. (2007). Fatal attraction: Salience, naïveté, and sophistication in experimental "hide-and-seek" games. *The American Economic Review*, 97(5):1731–1750.

van Damme, E. (1989). Stable equilibria and forward induction. *Journal of Economic Theory*, 48:476–469.

Dessalles, J.-L. (1998). Altruism, status, and the origin of relevance. In Hurford, J. R., Studdert-Kennedy, M., and Knight, C., editors, *Approaches to the Evolution of Language: Social and Cognitive Bases*, pages 130–147. Cambridge University Press.

Ettinger, D. and Jehiel, P. (2008). A theory of deception. Unpublished Manuscript.

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

Farrell, J. and Rabin, M. (1996). Cheap talk. *The Journal of Economic Perspectives*, 10(3):103–118.

Franke, M. (2008). Meaning and inference in case of conflict. In Balogh, K., editor, *Proceedings of the 13th ESSLLI Student Session*, pages 65–74.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, 1:60–79.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, 144:517–546.

Green, M. S. (1995). Quantity, volubility, and some varieties of discourse. *Linguistics and Philosophy*, 18:83–112.

Grice, P. H. (1989). *Studies in the Ways of Words*. Harvard University Press.

Groenendijk, J. and Stokhof, M. (1984). *Studies in the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, Universiteit van Amsterdam.

Hamilton, W. D. (1963). The evolution of altruistic behavior. *American Naturalist*, 97:354–356.

Heap, S. P. H. and Varoufakis, Y. (2004). *Game Theory - A Critical Text (Second Edition)*. Routledge.

Hirschberg, J. (1985). *A theory of scalar implicature*. PhD thesis, University of Pennsylvania.

Ho, T.-H., Camerer, C., and Weigelt, K. (1998). Iterated dominance and iterated best response in experimental “p-beauty contests”. *The American Economic Review*, 88(4):947–969.

Horn, L. R. (1984). Towards a new taxonomy for pragmatic inference: Q-based and I-based implicatures. In Shiffrin, D., editor, *Meaning, Form, and Use in Context*, pages 11–42. Georgetown University Press, Washington.

Jäger, G. (2007). Game dynamics connects semantics and pragmatics. In Pietarinen, A.-V., editor, *Game Theory and Linguistic Meaning*, pages 89–102. Elsevier.

Jäger, G. (2008). Game theory in semantics and pragmatics. Manuscript, University of Bielefeld.

Kohlberg, E. and Mertens, J.-F. (1986). On the strategic stability of equilibria. *Econometrica*, 54(5):1003–1037.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press, Cambridge, UK.

Lewis, D. K. (1969). *Convention.* Harvard University Press, Cambridge.

McAndrew, F. (2002). New evolutionary perspectives on altruism: Multilevel selection and costly-signaling theories. *Current directions in Psychological Science*, 11(2):79–82.

Merin, A. (1999). Information, relevance, and social decisionmaking: Some principles and results of decision-theoretic semantics. In *Logic, Language, and Computation*, volume II, pages 179–221. CSLI Publications.

Milgrom, P. and Roberts, J. (1986). Relying on the information of interested parties. *RAND Journal of Economics*, 17(1):18–32.

Parikh, P. (2001). *The Use of Language.* CSLI Publications, Stanford, California.

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52(4):1029–1050.

Rabin, M. (1990). Communication between rational agents. *Journal of Economic Theory*, 51:144–170.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5):1281–1302.

van Rooij, R. and Schulz, K. (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy*, 29:205–250.

Rubinstein, A., Tversky, A., and Heller, D. (1996). Naïve strategies in competitive games. In Albers, W., Güth, W., Hammerstein, P., Moldovanu, B., and van Damme, E., editors, *Understanding Strategic Interaction – Essays in Honor of Reinhard Selten*, pages 394–102. Springer Verlag, Berlin.

Sally, D. (2003). Risky speech: Behavioral game theory and pragmatics. *Journal of Pragmatics*, 35:1223–1245.

Schelling, T. C. (1960). *The Strategy of Conflict.* Harvard University Press, Cambridge, Massachusetts.

Selten, R. (1975). A reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4:25–55.

Shimoji, M. (2002). On forward induction in money burning games. *Economic Theory*, 19:637–648.

Solan, L. M. and Tiersma, P. M. (2005). *Speaking of Crime.* The Chicago Series in Law and Society. University of Chicago Press, Chicago.

Spence, A. M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87:355–374.

Sperber, D. and Wilson, D. (1995). *Relevance: Communication and Cognition (2nd ed.)*. Blackwell, Oxford.

Sperber, D. and Wilson, D. (2004). Relevance theory. In Horn, L. R. and Ward, G., editors, *Handbook of Pragmatics*, pages 607–632. Blackwell, Oxford.

Stahl, D. O. and Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10:218–254.

Stalnaker, R. (1998). Belief revision in games: Forward and backward induction. *Mathematical Social Sciences*, 36:31–56.

Stalnaker, R. (2006). Saying and meaning, cheap talk and credibility. In Benz et al. (2006), pages 83–100.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46:35–57.

Zahavi, A. (1975). Mate selection - a selection for a handicap. *Journal of Theoretical Biology*, 53:205–214.

Zahavi, A. (1990). Arabian babblers: the quest for social status in a cooperative breeder. In Stacey, P. and Koenig, W. D., editors, *Cooperative Breeding in Birds: Long Term Studies of Ecology and Behaviour*, pages 103–130. Cambridge University Press.

Zahavi, A. and Zahavi, A. (1997). *The Handicap Principle*. Oxford University Press.