MEANING & INFERENCE IN CASE OF CONFLICT

Michael Franke
Universiteit van Amsterdam, ILLC
M.Franke@uva.nl

Abstract. This paper applies a model of boundedly rational "level-k thinking" (c.f. Stahl & Wilson 1995, Crawford 2003, Camerer, et al. 2004) to a classical concern of game theory: when is information credible and what shall I do with it if it is not? The model presented here extends and generalizes recent work in game-theoretic pragmatics (Stalnaker 2006, Jäger 2007, Benz & van Rooij 2007). Pragmatic inference is modeled as a sequence of iterated best responses, defined here in terms of the interlocutors' epistemic states. Credibility considerations are a special case of a more general pragmatic inference procedure at each iteration step. The resulting analysis of message credibility improves on previous game-theoretic analyses, is more general and places credibility in the linguistic context where it, arguably, belongs.

1. Semantic Meaning and Credible Information in Signaling Games

The perhaps simplest game-theoretic model of language use is a *signaling game with meaningful signals*. A sender S observes the state of the world $t \in T$ in private and chooses a message m from a set of alternatives M all of which are assumed to be meaningful in the (unique and commonly known) language shared by S and a receiver R. In turn, R observes the sent message and chooses an action a from a given set A. In general, the payoffs for both S and R depend on the state t, the sent message m and the action a chosen by the receiver. Formally, a SIGNALING GAME WITH MEANINGFUL SIGNALS is a tuple $\langle \{S, R\}, T, \Pr, M, [\![\cdot]\!], A, \mathrm{U}_S, \mathrm{U}_R \rangle$ where $\Pr \in \Delta(T)$ is a probability distribution over T; $[\![\cdot]\!] : M \rightarrow \mathcal{P}(T)$ is a semantic denotation function and $\mathrm{U}_{S,R} : M \times A \times T \rightarrow \mathbb{R}$ are utility functions for both sender and receiver.[1] We can conceive of such signaling games as abstract mathematical models of a conversational context whose most important features they represent: the interlocutors' beliefs, behavioral possibilities and preferences.

If a signaling game is a context model, the game's *solution concept* is what yields a prediction of the behavior of agents in the modelled conversational situation. The following easy example of a *scalar implicature*, e.g., the inference that *not all* students came when hearing the sentence "Some of the students came", makes this distinction clear. A simple context model for this case is the signaling game G1:[2] there are two states $t_{\exists\neg\forall}$ and $t_{\forall}$, two messages $m_{\texttt{some}}$ and $m_{\texttt{all}}$ with semantic meaning as indicated and two receiver interpretation actions $a_{\exists\neg\forall}$ or $a_{\forall}$ which correspond one-to-one with the states; sender and receiver payoffs are aligned: an implementation of the standard assumption that conversation and implicature calculation revolve around the *cooperative principle* (Grice 1989). A

[1] I will assume throughout that (i) all sets T, M and A are non-empty and finite, that (ii) $\Pr(t) > 0$ for all $t \in T$, that (iii) for each state t there is at least one message m which is true in that state and that (iv) no message is contradictory, i.e., there is no m for which $[\![m]\!] = \emptyset$.

[2] Unless indicated, I assume that states are equiprobable in example games.

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

G1: "Scalar Implicatures"

	$a_{\texttt{mate}}$	$a_{\texttt{ignore}}$	$m_{\texttt{high}}$	$m_{\texttt{low}}$
$t_{\texttt{high}}$	1,1	0,0	$\surd$	–
$t_{\texttt{low}}$	1,0	0,1	–	$\surd$

G2: "Partial Conflict"

solution concept, whatever it may be, should then ideally predict that $S^{t_\forall}$ ($S^{t_{\exists\neg\forall}}$) chooses $m_{\texttt{some}}$ ($m_{\texttt{all}}$) and the receiver responds with action $a_{\exists\neg\forall}$ ($a_\forall$).[3]

It is obvious that in order to arrive at this prediction, a special role has to be assigned to the conventional, semantic meaning of the messages involved. For instance, in the above example *anti-semantic* play, as we could call it, that simply reverses the use of messages, should be excluded. Most game-theoretic models of language use hard-wire semantic meaning into the game play, either as a restriction on available moves of sender and receiver, or into the payoffs, but in both cases effectively enforcing truthfulness and trust. This is fine as long as conversation is mainly cooperative and preferences aligned. But let's face it: the central Gricean assumption of cooperation is an optimistic idealization after all; conflict, lies and deceit are as ubiquitous as air. But then, hard-wiring of truthfulness and trust limits the applicability of our models as it excludes the possibility that senders may wish to *mislead* their audience. We should aim for more general models and, ideally, let the agents, not the modeller decide when to be truthful and what to trust.

Opposed to hard-wiring truthfulness and trust, the most liberal case at the other end of the spectrum is to model communication, not considering reputation or further psychological constraints at all, as *cheap talk*. Here messages do not impose restrictions on the game play and are entirely payoff irrelevant: $\mathrm{U}_{S,R}(m, a, t) = \mathrm{U}_{S,R}(m', a, t)$ for all $m, m' \in M$, $a \in A$ and $t \in T$. However, if talk is cheap, yet exogenously meaningful, the question arises how to integrate semantic meaning into the game. Standard solution concepts, such as *sequential equilibrium* or *rationalizability*, are too weak to predict anything reasonable in this case: they allow for nearly all anti-semantic play and also for *babbling*, where signals are sent, as it were, arbitrarily and therefore ignored by the receiver.

In response to this problem, game theorists have proposed various refinements of the standard solution concepts based on the notion of *credibility*.[4] The idea is that semantic meaning should be respected (in the solution concept) wherever this is reasonable in view of the possibly diverging preferences of interlocutors. As an easy example, look at game G2 where S is of either a high quality or a low quality type, and where R would like to pair with $S^{t_{\texttt{high}}}$ only, while S wants to pair with R irrespective of her type. Interests are in partial conflict here and, intuitively, a costless, non-committing message $m_{\texttt{high}}$ is *not* credible, because $S^{t_{\texttt{low}}}$ would have all reason to send it untruthfully. Therefore, intuitively, R should ignore whatever S says in this game. In general, if nothing prevents S from babbling, lying or deceiving, she might as well do so; whenever she even has an incentive to, she certainly will. For the receiver the central question becomes: when is a signal credible and what should I do if it is not?

This paper offers a fresh look at this classical problem of game theory. The novelty is, so to speak, a "linguistic turn": I suggest that credibility considerations are *pragmatic inferences*, in some sense very much alike —and in another sense very much unlike— con-

[3] For $t \in T$, I write S^t as an abbreviation for "a sender of type t".

[4] The standards in the debate about credibility were set by Farrell (1993) for equilibrium and by Rabin (1990) for rationalizability. I will mainly focus on these two classical papers here for reasons of space.

versational implicatures. I argue that this linguistic approach to credibility of information improves on the classical game-theoretic analyses by Farrell (1993) and Rabin (1990). In order to implement conventional meaning of signals in a cheap talk model, the present paper takes an *epistemic approach* to the solution of games: the model presented in this paper spells out the reasoning of interlocutors in terms of their beliefs about the behavior of their opponents as *a sequence of iterated best responses* (IBR) which takes semantic meaning as a starting point. For clarity: the IBR model places no restriction whatsoever on the use of signals; conventional meaning is implemented merely as a focal element in the deliberation of agents. This way, the IBR model extends recent work in game-theoretic pragmatics (Jäger 2007, Benz & van Rooij 2007), to which it adds generality by taking diverging preferences into account and by implementing the basic assumptions of "level-k models" of reasoning in games (cf. Stahl & Wilson 1995, Crawford 2003, Camerer et al. 2004). In particular, agents in the model are assumed to be *boundedly rational* in the sense that each agent computes only finitely many steps of the best response sequence. Section 2. scrutinizes the notion of credibility, section 3. spells out the formal model and section 4. discusses its properties and predictions.

2. Credibility and Pragmatic Inference

The classical idea of message credibility is due to Farrell (1993). Farrell seeks an equilibrium refinement that pays due respect to the semantic meaning of messages. His notion of credibility is therefore tied to a given reference equilibrium as a status quo. According to Farrell, then, a message m is FARRELL-CREDIBLE with respect to a given equilibrium if all $t \in [\![m]\!]$ prefer the receiver to interpret m literally, i.e., to play a best response to the belief $\Pr(\cdot \mid [\![m]\!])$ that m is true, over the equilibrium play, while no type $t \notin [\![m]\!]$ does.

A number of objections can be raised against Farrell-credibility. First of all, the definition requires *all* types in $[\![m]\!]$ to prefer a literal interpretation of m over the reference equilibrium. This makes sense, under Farrell's *Rich Language Assumption* (RLA) that for every $X \subseteq T$ there is a message m with $[\![m]\!] = X$. This assumption is prevalent in game-theoretic discussions of credibility, but restricts applicability. I will show in section 4. that this assumption seriously restricts Rabin's (1990) account. But for now, suffice it to say that, in particular, the RLA excludes models like G1, used to study pragmatic inference in the light of (partial) inexpressibility. I will drop the RLA here to aim for more generality and compatibility with linguistic pragmatics.[5] Doing so, implies amending Farrell-credibility to require only that *some* types in $[\![m]\!]$ prefer a literal interpretation of m over the reference equilibrium.

Still, there are further problems. Matthews, et al. (1991) criticize Farrell-credibility as being *too strong*. Their argument builds on example G3. Compared to the babbling equilibrium, in which R performs a_3, messages m_1 and m_2 are intuitively credible: both S^{t_1}, as well as S^{t_2} have good reason to send m_1 and m_2 respectively. Communication seems possible and utterly plausible. However, neither message is Farrell-credible, because for

[5] A reviewer points out that the RLA has a correspondent in the linguistic world in Katz's (1981) "principle of effability". The reviewer supports dropping the RLA, because otherwise pragmatic inferences is limited to context and effort considerations. It is also very common (and, to my mind, reasonable) to restrict attention to certain alternative expressions only, namely those that are salient (in context) after observing a message. Of course, game theory is silent as to where the alternatives come from, since this is a question for the linguist, perhaps even the syntactician (cf. Katzir 2007).

	a_1	a_2	a_3	m_1	m_2
t_1	4,3	3,0	1,2	√	–
t_2	3,0	4,3	1,2	–	√

G3: “Best Message Counts”

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

G4: “Further Iteration”

$i, j \in \{1, 2\}$ and $i \neq j$ not only S^{t_j}, but also S^{t_i} prefers R to play a best response to a literal interpretation of m_j, which would trigger action a_j, over the no-communication outcome a_3. The problem with Farrell's notion is obviously that just doing better than equilibrium is not enough reason to send a message, when sending another message is *even better* for the sender. When evaluating the credibility of a message m, we have to take into account *alternative forms* that $t \notin [\![m]\!]$ might want to send.

Compare this with the scalar implicature in G1. Message $m_{\texttt{some}}$ is interpreted as communicating that the true state of affairs is $t_{\exists\neg\forall}$, because in $t_{\forall}$ the sender would have used $m_{\texttt{all}}$. In other words, the receiver discards a state $t \in [\![m]\!]$ as a possible sender of m because that type has a better message to send. Of course, such *pragmatic enrichment* does not make a message intuitively incredible, as it is still used in line with its semantic meaning. Intuitively speaking, in G1 S even *wants* R to draw this pragmatic inference.

This is, of course, different in G2. In general, if S wants to mislead, she intuitively wants the receiver to adopt a certain belief, but she does *not* want the receiver to realize that this belief might be false: we could say, somewhat loosely, that S wants her purported communicative intention to be recognized (and acted upon), but she does *not* want her *deceptive* intention to be recognized. Nevertheless, if the receiver does manage to recognize a deceptive intention, this too may lead to some kind of pragmatic inference, albeit one that the sender did not intend the receiver to draw. While the implicature in G1 rules *out* a semantically feasible possibility, credibility considerations, in a sense, do the exact opposite: message $m_{\texttt{high}}$ is *pragmatically weakened* in G2 by ruling *in* state $t_{\texttt{low}}$.

Despite the differences, there is a common core to both implicature and credibility inference. Here and there, the receiver seems to reason: which types of senders would send this message given that I believe it literally? Indeed, exactly this kind of reasoning underlies Benz & van Rooij's (2007) model of implicature calculation for the purely cooperative case. The driving observation of this paper is that the same reasoning might not only rule out states $t \in [\![m]\!]$ to yield implicatures but may also rule in states $t \notin [\![m]\!]$. When the latter is the case, m seems intuitively incredible. Still, the reasoning pattern by which implicatures and credibility-based inferences are computed is the same. On superficial reading, this view on message credibility can be found in Stalnaker (2006) :[6] call a message m BvRS-credible (Benz, van Rooij, Stalnaker) iff for some types $t \in [\![m]\!]$, but for no type $t \notin [\![m]\!]$ S^t's expected utility of sending m given that R interprets literally is at least as great as S^t's expected utility of sending any alternative message m'.

The notion of BvRS-credibility matches our intuitions in all the cases discussed so far, but it is, in a sense, self-refuting, as G4 from Matthews et al. (1991) shows. In this game,

[6] It is unfortunately not entirely clear to me what exactly Stalnaker's proposal amounts to, as insightful as it might be, because the account is not fully spelled out formally. The basic idea seems to be that (something like) the notion of BvRS-credibility, as it is called here, should be integrated as a constraint on receiver beliefs —believe a message iff it is BvRS-credible— into an epistemic model of the game together with some appropriate assumption of (common) belief in rationality. The class of game models that satisfies rationality and credibility constraints would then ultimately define how signals are used and interpreted.

all the available messages m_{12}, m_{23} and m_{13} are BvRS-credible, because if R interprets literally S^{t_1} will use message m_{12}, S^{t_2} will use message m_{23} and S^{t_3} will use message m_{13}. No message is used untruthfully by any type. However, if R realizes that exactly S^{t_1} uses message m_{12}, he would rather not play a_2, but a_1. But if the sender realizes that message m_{12} triggers the receiver to play a_1, suddenly S^{t_3} wants to send m_{12} *untruthfully*. This example shows that BvRS-credibility is a reliable start, but stops too short. If messages are deemed credible and therefore believed, this may create an incentive to mislead. What seems needed to rectify the formal analysis of message credibility is a fully spelled-out model of iterated best responses that starts in the Benz-van-Rooij-Stalnaker way and then carries on iterating. Here is such a model.

3. The IBR Model and its Assumptions

3.1. Assumptions: Focal Meaning and Bounded Rationality

The IBR model presented in this paper rests on three assumptions with which it also sets itself apart from previous best-response models in formal pragmatics (Jäger 2007, Benz & van Rooij 2007, Jäger 2008). The first assumption is the *Focal Meaning Assumption*: semantic meaning is focal in the sense that the sequence of best responses starts with a purely semantic truth-only sender strategy. Semantic meaning is also assumed focal in the sense that throughout the IBR sequence R believes messages to be truthful unless S has a positive incentive to be untruthful. This is the second, so called *Truth Ceteris Paribus Assumption* (TCP). These two (epistemic) assumptions assign semantic meaning its proper place in this model of cheap-talk communication.

The third assumption is the *Bounded Rationality Assumption*: I assume that players in the game have limited resources which allow them to reason only up to some finite iteration depth k. At the same time I take agents to be *overconfident*: each agent beliefs that she is smarter than her opponent. Camerer et al. (2004) make an empirical case for these assumptions about the psychology of reasoners.[7] However, for simplicity, I do not implement Camerer et al.'s (2004) *Cognitive Hierarchy Model* in full. Camerer et al. assume that each agent who is able to reason up to strategic depth k has a proper belief about the population distribution of players who reason up to depth $l < k$, but I will assume here, just to keep things simple, that each player believes that she is exactly one step ahead of her opponent (cf. Crawford 2003, Crawford 2007). (I will discuss this simplifying assumption critically in section 4..)

3.2. Beliefs & Best Responses

Given a signaling game, a SENDER SIGNALING-STRATEGY is a function $\sigma \in \mathcal{S} = (\Delta(M))^T$ and a RECEIVER RESPONSE-STRATEGY is a function $\rho \in \mathcal{R} = (\Delta(A))^M$. In order to define which strategies are best responses to a given belief, we need to define the game-relevant beliefs of both S and R. Since the only uncertainty of S concerns what

[7] A good intuitively accessible example why this should be is a so-called *beauty contest game* (cf. Ho, et al. 1998). Each player from a group of size $n > 2$ chooses a number from 0 to 100. The player closest to $2/3$ the average wins. When this game is played with a group of subjects who have never played the game before, the usual group average lies somewhere between 20 to 30. This is quite far from the group average 0 which we would expect from common (true) belief in rationality. Everybody seems to believe that they are just a smarter than everybody else, without noticing their own limitations.

R will do, the set of relevant SENDER BELIEFS Π_S is just the set of receiver response-strategies: $\Pi_S = \mathcal{R}$. On the receiver's side, we may say, with some redundancy, that there are three components in any game-relevant belief (cf. Battigalli 2006): firstly, R has a *prior* belief $\Pr(\cdot)$ about the true state of the world; secondly, he has a belief about the sender's signaling strategy; and thirdly, he has a *posterior* belief about the true state *after* hearing a message. Posteriors should be derived by Bayesian update from the former two components, but also specify R's beliefs after unexpected *surprise messages*. Taken together, the set of relevant RECEIVER BELIEFS Π_R is the set of all triples $\langle \pi^1_R, \pi^2_R, \pi^3_R \rangle$ for which $\pi^1_R = \Pr$, $\pi^2_R \in \mathcal{S} = (\Delta(M))^T$ and $\pi^3_R \in (\Delta(T))^M$ such that for any $t \in T$ and $m \in M$ if $\pi^2_R(t, m) \neq 0$, then:

$$\pi^3_R(m, t) = \frac{\pi^1_R(t) \times \pi^2_R(t, m)}{\sum_{t' \in T} \pi^1_R(t') \times \pi^2_R(t', m)}.$$

Given a sender belief $\rho \in \Pi_S$, say that σ is a BEST RESPONSE SIGNALING STRATEGY to belief ρ iff for all $t \in T$ and $m \in M$ we have:

$$\sigma(t, m) \neq 0 \rightarrow m \in \arg\max_{m' \in M} \sum_{a \in A} \rho_{m'}(a) \times \mathrm{U}_S(m', a, t)$$

The set of all such best responses to belief ρ is denoted by $\mathcal{S}(\rho)$. Given a receiver belief $\pi_R \in \Pi_R$ say that ρ is a BEST RESPONSE STRATEGY to belief π_R iff for all $m \in M$ and $a \in A$ we have:

$$\rho(m, a) \neq 0 \rightarrow a \in \arg\max_{a' \in A} \sum_{t \in T} \pi^3_R(m, t) \times \mathrm{U}_R(m, a', t)$$

The set of all such best responses to belief π_R is denoted by $\mathcal{R}(\pi_R)$. Also, if $\Pi'_R \subseteq \Pi_R$ is a set of receiver beliefs, let $\mathcal{R}(\Pi'_R) = \bigcup_{\pi_R \in \Pi'_R} \mathcal{R}(\pi_R)$.

3.3. Strategic Types and the IBR sequence

In line with the Bounded Rationality Assumption of Section 3.1., I assume that senders and receivers are of different *strategic types*. Strategic types correspond to the level k of strategic depth a player in the game performs (while believing she thereby outperfoms her opponent by exactly one step of reasoning). I will give an inductive definition of strategic types in terms of players beliefs, starting with a fixed strategy σ^*_0 of S_0.[8] Then, for any $k \geq 0$, R_k is characterized by a belief set $\pi^*_{R_k} \subseteq \Pi_R$ that S is a level-k sender and S_{k+1} is characterized by a belief $\pi^*_{S_{k+1}} \in \Pi_S$ that R is a level-k receiver.

I assume that S_0 plays according to the signaling strategy σ^*_0 which simply sends any true message with equal probability in all states. There need not be any belief to which this is a best response, as level-0 senders are (possibly irrational) dummies to implement the Focal Meaning Assumption. R_0 then believes that he is facing S_0. With unique σ^*_0, which sends all messages in M with positive probability (M is finite and contains no contradictions), R_0 is characterized entirely by the unique belief $\pi^*_{R_o}$ that S plays σ^*_0.

In general, R_k believes that he is facing a level-k sender. For $k > 0$, S_k is characterized by a belief $\pi^*_{S_k} \in \Pi_S$. R_k consequently believes that S_k plays a best response $\sigma_k \in$

[8] I will write S_k and R_k to refer to a sender or receiver of strategic type k. Likewise, S^t_k refers to a sender of strategic type k and knowledge type t.

$\mathcal{S}(\pi^*_{S_k})$ to this belief. We can leave this unrestricted and assume that R_k considers any $\sigma_k \in \mathcal{S}(\pi^*_{S_k})$ possible. But it will transpire that for an intuitively appealing analysis of message credibility we need to assume that R_k takes S_k to be truthful all else being equal (see also discussion in section 4.). We implement the TCP Assumption of Section 3.1. as a restriction $\mathcal{S}^*(\pi^*_{S_k}) \subseteq \mathcal{S}(\pi^*_{S_k})$ on signaling strategies held possible by R. Of course, even when restricted, there need not be a unique signaling strategy here. As a general tie-break rule, assume the "principle of insufficient reason" that all $\sigma_k \in \mathcal{S}^*(\pi^*_{S_k})$ are equiprobable to R_k. That means that R_k effectively believes that his opponent is playing response strategy

$$\sigma^*_k(t,m) = \frac{\sum_{\sigma \in \mathcal{S}^*(\pi^*_{S_k})} \sigma(t,m)}{|\mathcal{S}^*(\pi^*_{S_k})|}.$$

This fixes R_k's beliefs about the behavior of his opponent, but it need not fix R_k's belief π^3_R about surprise messages. Since this matter is intricate and moreover R_k's *counterfactual beliefs* do not play a crucial role in any examples discussed in this paper, I will not pursue this issue at all in this paper (but see also footnote 10 below). In general, let us say that R_k is characterized by any belief whose second component is σ^*_k and whose third component satisfies *some* (coherent, but possibly vacuous) assumption about the interpretation of surprise messages. Let, $\pi^*_{R_k} \subseteq \Pi_R$ be the set of all such beliefs. R_k is then fully characterized by $\pi^*_{R_k}$.

In turn, S_{k+1} believes that her opponent is a level-k receiver who plays a best response $\rho_k \in \mathcal{R}(\pi^*_{R_k})$. With the above tie-break rule S_{k+1} is fully characterized by the belief

$$\rho^*_k(m,a) = \frac{\sum_{\rho \in \mathcal{R}(\pi^*_{R_k})} \rho(m,a)}{|\mathcal{R}(\pi^*_{R_k})|}.$$

3.4. Credibility and Inference

Define that a signal m is k-OPTIMAL in t iff $\sigma^*_{k+1}(t,m) \neq 0$. The set of k-optimal messages in t are all messages that R_{k+1} believes S^t_{k+1} might send (thus taking the TCP Assumption into account).[9] Similarly, distill from R's beliefs his INTERPRETATION-STRATEGY $\delta : M \to \mathcal{P}(T)$ as given by belief π_R: $\delta_{\pi_R}(m) = \{t \in T \mid \pi^3_R(m,t) \neq 0\}$. This simply is the *support* of the posterior beliefs of R after receiving message m. Let's write δ_k for the interpretation strategy of a level-k receiver.

For any $k > 0$, since S_k believes to face R_{k-1} with interpretation strategy δ_{k-1}, wanting to send message m would intuitively count as an attempt to *mislead* if sent by S^t_k just in case $t \notin \delta_{k-1}(m)$. Such an attempt would moreover be untruthful if $t \notin [\![m]\!]$. While R_{k-1} would be deceived, R_k would see through the attempted deception. From R_k's point of view, who adheres to the TCP Assumption, a message m is incredible if it is $k-1$-optimal in some $t \notin [\![m]\!]$. But then R_k will include t in his interpretation of m: recognizing a deceptive intention leads to pragmatic inference. In general, we should consider a message m credible unless some type $t \notin [\![m]\!]$ would want to use m somewhere along the IBR sequence; precisely, m is CREDIBLE iff $\delta_k(m) \subseteq [\![m]\!]$ for all $k \geq 0$.[10]

[9] Without the TCP Assumption, 0-optimality would be equivalent to the notion of an *optimal assertion* in Benz & van Rooij (2007).

[10] It may seem that messages which would not be sent by any type (after the first round or later) come out credible under this definition, which would not be a good prediction. (Thanks to Daniel Rothschild (p.c.) for

	a_1	a_2	m_{12}	m_3
t_1	1,1	0,0	√	–
t_2	0,0	1,1	√	–
t_3	0,0	1,1	-	√

G5: “White Lie”

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

G6: “Some Game without a Name”

4. Discussion

The IBR model makes intuitively correct predictions about message credibility for the games considered so far. In G1, R_0 responds to $m_{\texttt{some}}$ with the appropriate action $a_{\exists\neg\forall}$, but still interprets $\delta_0(m_{\texttt{some}}) = \{t_{\exists\neg\forall}, t_\forall\}$. In turn, R_1 interprets as $\delta_1(m_{\texttt{some}}) = \{t_{\exists\neg\forall}\}$; he has pragmatically enriched the semantic meaning by taking the sender’s payoff structure and available messages into account. After one round a fixed-point is reached, with fully revealing credible signaling in accordance with intuition. In G2, IBR predicts that both $S_1^{t_{\texttt{high}}}$ and $S_1^{t_{\texttt{low}}}$ will use $m_{\texttt{high}}$ which is therefore not credible. In G3, also fully revealing communication is predicted and for G4 IBR predicts that all messages are credible for R_0 and R_1, but not for R_2, hence incredible as such. In general, the IBR model predicts that communication in games of pure coordination is always credible:

Proposition 4..1. Take a signaling game with $T = A$ and $\mathrm{U}_{S,R}(\cdot, t, t') = c > 0$ if $t = t'$ and 0 otherwise. Then $\delta_k(m) \subseteq [\![m]\!]$ for all k and m.

Proof. Clearly, $\delta_0(m) \subseteq [\![m]\!]$ for arbitrary m. So assume that $\delta_k(m) \subseteq [\![m]\!]$. In this case S_{k+1}^t will use m only if $t \in \delta_k(m)$. But then $t \in [\![m]\!]$ and therefore $\delta_{k+1}(m) \subseteq [\![m]\!]$. □

However, the IBR model does *not* guarantee generally that communication is credible even when preferences are *perfectly aligned*, i.e., $\mathrm{U}_S = \mathrm{U}_R$. This may seem surprising at first, but is due naturally to the possibility of, what we could call, *white lies*: untruthful signaling that is beneficial for the receiver. These may occur if the set of available signals is not expressive enough. As an easy example, consider G5 where S^{t_2} will use m_3 untruthfully to induce action a_2, which, however, is best for both receiver and sender.

To understand the central role of the TCP assumption in the present proposal, consider the game G6. In G6, R_0 has the following posterior beliefs: after hearing message m_{12} he rules out t_3 and believes that t_2 is three times as likely as t_1; similarly, after hearing message m_{23} he rules out t_1 and believes that t_2 is three times as likely as t_3. Consequently, R_0 responds to both signals with a_2. Now, $S_1^{t_1}$, for instance, does not care which message to choose from, as far as her expected utilities are concerned. But R_1 nevertheless assumes that $S_1^{t_1}$ speaks truthfully. It’s thanks to the TCP Assumption that IBR predicts messages to be credible in this game.

G6 also shows a difference between the IBR model and Rabin’s (1990) model of credible communication, which superficially look very similar. Rabin’s model consists of two components: the first component is a definition of message credibility which is *almost* a

pointing this out to me.) However, this is not quite right: we get into this predicament only for *some* versions of the IBR sequence, not for others. It all depends on how the receiver forms his counterfactual beliefs. If, for instance, we assume that R *rationalizes* observed behavior even if it surprises him, we can keep the definition unchanged: if no type whatsoever has an outstanding reason to send m, the receiver’s posterior beliefs after m will support any type. So, unless m is tautologous, it is incredible. Still, Rothschild’s criticism is appropriate: the definition of message credibility offered here is, in a sense, incomplete as long as we do not properly define the receiver’s counterfactual beliefs; something left for another occasion.

two-step iteration of best responses starting from the semantic meaning; the second component is iterated strict dominance around a fixed core set of Rabin-credible messages being sent truthfully and believed. In particular, Rabin requires for m to be credible that m induces, when taken literally, exactly the set of all sender-best actions (from the set of actions that are inducible by *some* receiver belief) of all $t \in [\![m]\!]$. This is defensible under the Rich Language Assumption, but both messages in G6 fail this requirement. Consequently, with no credible message to restrict iterated strict dominance, Rabin's model predicts a total anything-goes for game G6. This shows the limited applicability of approaches to message credibility that are inseparable from the Rich Language Assumption. The present notion of message credibility and the IBR model are not restricted in this sense and fare well with (partial) inexpressibility and the resulting inferences.

To wrap up: as a solution concept, the epistemic IBR model offers, basically, a set of beliefs, viz., beliefs obtained under certain assumptions about the psychology of agents from a sequence of iterated best responses. I do not claim that this model is a reasonable model for human reasoning in general. Certainly, the simplifying assumption that players believe that they are facing a level-k opponent, and not possibly a level-$l < k$ opponent, is highly implausible proportional to k, but especially so for agents that have, in a manner of speaking, already reasoned themselves through a circle multiple times. (It is easily verified that for finite M and T the IBR sequence always enters a circle after some $k \in \mathbb{N}$.)[11] Still, I wish to defend that the IBR model *does* capture (our intuitions about) certain aspects of (idealized) linguistic behavior, namely pragmatic inference in cooperative and non-cooperative situations. Whether it is a plausible model of belief formation and reasoning in the envisaged linguistic situations is ultimately an empirical question.

In conclusion, the IBR model offers a novel perspective on message credibility and the pragmatic inferences based on this notion. The model generalizes existing game-theoretical models of pragmatic inference by taking conflicting interests into account. It also generalizes game-theoretic accounts of credibility by giving up the Rich Language Assumption. The explicitly epistemic perspective on agents' deliberation assigns a natural place to semantic meaning in cheap-talk signaling games as a focal starting point. It also highlights the unity in pragmatic inference: in this model both credibility-based inferences and implicatures are different outcomes of the same reasoning process.

Acknowledgements

I'd like to thank Tikitu de Jager, Robert van Rooij, Daniel Rothschild, Marc Staudacher and three anonymous referees for insightful comments, help and discussion. I moreover benefited greatly from discussing with Gerhard Jäger an early version of his paper (Jäger 2008), which also defines and applies a general iterated best response model different from what I did here. Also, I am thankful to Sven Lauer for waking my interest by first explaining to me with enormous patience some puzzles about credibility that I did not fully understand at the time (see Lauer 2007). Errors are my own.

[11] It is tempting to assume that "looping reasoners" may have an *Aha-Erlebnis* and to extend the IBR sequence by transfinite induction assuming, for instance, that level-ω players best respond to the belief that the IBR sequence is circling. I do not know whether this is necessary and/or desirable for linguistic applications. We should keep in mind though that in some cases human reasoners may not get to the ideal level of reasoning in this model and in others they might even go beyond it.

References

P. Battigalli (2006). ‘Rationalization in Signaling Games: Theory and Applications’. *International Game Theory Review* **8**(1):67–93.

A. Benz & R. van Rooij (2007). ‘Optimal Assertions and what they Implicate’. *Topoi* **26**:63–78.

C. F. Camerer, et al. (2004). ‘A Cognitive Hierarchy Model of Games’. *The Quarterly Journal of Economics* **119**(3):861–898.

V. P. Crawford (2003). ‘Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions’. *American Economic Review* **93**(1):133–149.

V. P. Crawford (2007). ‘Let’s Talk It Over: Coordination Via Preplay Communication With Level-k Thinking’. Unpublished Manuscript.

J. Farrell (1993). ‘Meaning and Credibility in Cheap-Talk Games’. *Games and Economic Behavior* **5**:514–531.

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

T.-H. Ho, et al. (1998). ‘Iterated Dominance and Iterated Best Response in Experimental “p-Beauty Contests”’. *The American Economic Review* **88**(4):947–969.

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

G. Jäger (2008). ‘Game Theory in Semantics and Pragmatics’. Manuscript, University of Bielefeld.

J. J. Katz (1981). *Language and Other Abstract Objects*. Basil Blackwell.

R. Katzir (2007). ‘Structurally-Defined Alternatives’. To appear in Linguistics and Philosophy.

S. Lauer (2007). ‘Some kinds of deception do not occur: Credibility and the maxim of sincerity’. Unpublished Manuscript. Amsterdam, Stanford.

S. A. Matthews, et al. (1991). ‘Refining Cheap Talk Equilibria’. *Journal of Economic Theory* **55**:247–273.

M. Rabin (1990). ‘Communication between Rational Agents’. *Journal of Economic Theory* **51**:144–170.

D. O. Stahl & P. W. Wilson (1995). ‘On Players’ Models of Other Players: Theory and Experimental Evidence’. *Games and Economic Behavior* **10**:218–254.

R. Stalnaker (2006). ‘Saying and Meaning, Cheap Talk and Credibility’. In A. Benz, G. Jäger, & R. van Rooij (eds.), *Game Theory and Pragmatics*, pp. 83–100. Palgrave MacMillan.