Disambiguation Games in Extended and Strategic Form

Sascia Pavan
s.pavan@tele2.it

June 30, 2008

Abstract

The aim of this paper is to pursue the line of research initiated by Prashant Parikh which gives content and rigour to the intuitive idea that speaking a language is a rational activity. He employs the most promising tool to that end, namely game theory. I consider one of his examples as a sample case, and the model I build is a slight modification of that developed by him. I argue that my account has some advantage, yet many of the key ideas employed are left unchanged. I analyse this model in detail, describing some of its formal features. I conclude raising a problem that has not been analysed yet, sketching a plausible solution.

1 Introduction

The case I want to analyse concerns sentences like

(1) Every ten minutes a man gets mugged in New York.

This sentence has two readings, one is that there is a certain man in New York, either very unlucky, or reckless, or masochist, that is mugged every ten minutes. The other reading is that every ten minutes, some man or other, not necessarily the same, gets mugged in New York. Imagine an actual conversation where (1) is uttered, the problem is: How can the hearer guess the reading originally intended by the speaker? As for (1), we can hardly imagine a situation where the reading intended by the speaker is the first one – namely the unlucky, reckless, masochist interpretation – and where this is the reading selected by the hearer. An interesting feature of (1) is that one of the two possible readings entails the other, in this case the second reading is a logical consequence of the first. We can think of sentences sharing this same feature with (1), but such that they can be employed in a conversation where the intended reading is the logically stronger one. Consider

(2) All of my graduate students love a Finnish student in my Game-Theory class.

Suppose that (2) is uttered by a professor in Amsterdam. I do not know how many Finnish students studying game theory there are in Amsterdam. Assume there are very few of them. My intuition is that in most situations the hearer would infer that there is a unique Finnish student in the speaker's class that all graduate students love.

I will address the question how a speaker and a hearer can successfully communicate employing ambiguous sentences. I will use (1) and (2) as sample cases, because the logical feature they share – namely the fact that one of the two readings is a logical consequence of the other – imposes some constraints on the game-theoretic model that will be built that will greatly simplify the analysis. Yet, we can expect that the solution concepts proposed for these cases will at least provide hints for models covering a wider class of cases.

My starting point will be the account proposed by Prashant Parikh in several works (1992; 2001; 2006). The extensive form of his model can be represented by Figure 1. In his theory the speaker is player 1 and the hearer is player 2. As is customary in game theory, I will imagine that player 1 is male, and player 2 female. Player 2 has two options, she has to choose among two moves, namely the alternative interpretations A and B of some ambiguous sentence ϕ, and she does not know whether she is in the situation where player 1 means A or in the situation where he means B. Speaking technically, her *information set* contains two nodes labelled '2.c' and it is marked by an ellipsis. The root of the tree represents a chance event where 'Nature', determines whether 1 means A – let this be situation a – or means B – situation b. The prior probabilities of these alternatives are p and $1-p$, respectively, where $1 > p > 0$. If player 1 is in situation a, he can utter either ϕ or μ_a, and these two moves are labelled 'I' and 'E', respectively, where 'E' is short for 'explicit' and 'I' for 'implicit'. If he is in situation b, he can choose between ϕ and μ_b, and here the alternative moves are 'i' and 'e'. Since player 2 has a chance to move only if the game is in one of the states 2.c, when 1 chooses an unambiguous sentence, the game ends, otherwise 2 must choose among A and B, an then the game ends.

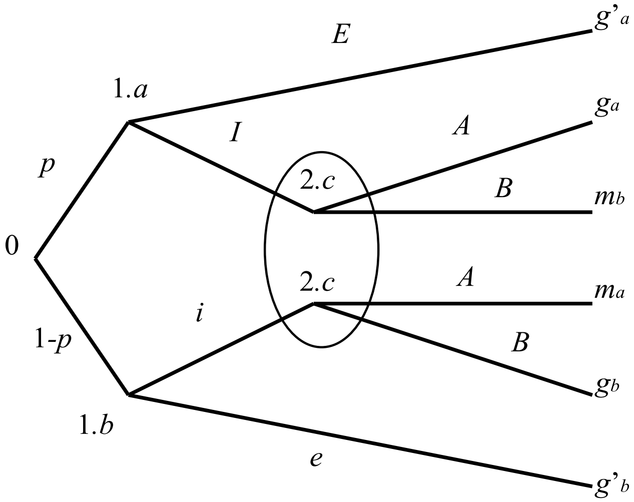

Figure 1: Disambiguation game: the extensive form

I am actually deviating from Parikh's original presentation. In Parikh's diagram there is a pair of trees, like in Figure 2, instead of a single tree, because he argues that player 2 'cannot construct anything' before 1's utterance (2001, p. 83), this is why he proposes the notion of a game of *partial information*. I prefer to stick to more traditional methods, since this new notion does not alter the relevant mathematical features of a game, it seems to be an unnecessary – but harmless – deviation from well-established standards. I observe that if Parikh is right in his claim that these disambiguation games should not be treated as ordinary games of imperfect information, the same would hold, for example, for Spence's 'model of education' (Spence, 1975), or the famous 'Beer or Quiche' (Cho and Kreps, 1987).

Parikh takes this to be a game of *pure coordination* (2001, pp. 29, 40n) where, at every terminal node, player 1's payoff is the same as player 2's. They get the lowest payoffs when the speaker chooses the ambiguous sentence and the hearer selects the wrong interpretation. The highest ones when he chooses the ambiguous sentence and she picks out the correct inter-

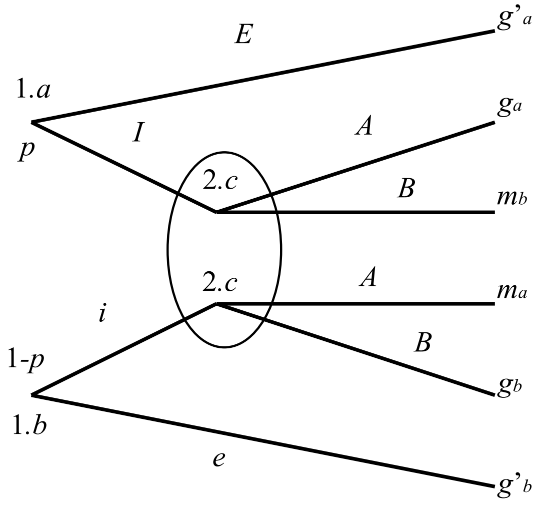

Figure 2: Parikh's game of *partial information*

pretation. The remaining payoffs are intermediate because they also amount to a successful communication but a less convenient one since the explicit sentence is longer than the ambiguous one. The normal representation of this game is the set $G = \{N, C_1, C_2, u\}$, where $N = \{1, 2\}$ is the set of players, $C_1 = \{Ei, Ee, Ii, Ie\}$ and $C_2 = \{A, B\}$ are the sets of their pure strategies, and u is their payoff function, hence a function from $C_1 \times C_2$ to the real line $\mathbb{R}$. It satisfies the pattern shown in Table 1. There are two Nash equilibria in pure strategies in the normal representation of the game, namely $(Ie], [A])$ and $([Ei], [B])$. None of the usual refinements can rule out any of these. Parikh's theory predicts that the players will tend to converge on the most efficient one. There are also infinitely many less efficient mixed equilibria.

2 Chance Events in Disambiguation Games

I argue that this model is wanting in one respect, namely the nature of the chance events, or, equivalently, of the private information possessed by the speaker. The speaker's intended meaning m is the speaker's intention to communicate that m, as such it is not much different from an intention to perform some other action whatsoever. In any case, we act upon our beliefs

	A	B
Ei	$p \times g'_a + (1-p) \times m_a$	$p \times g'_a + (1-p) \times g_b$
Ee	$p \times g'_a + (1-p) \times g'_b$	$p \times g'_a + (1-p) \times g'_b$
Ii	$p \times g_a + (1-p)m_a$	$p \times m_b + (1-p) \times g_b$
Ie	$p \times g_a + (1-p) \times g'_b$	$p \times m_b + (1-p) \times g'_b$

Table 1: Disambiguation game: the strategic form

and goals. Game theory models those situations where we must act upon our beliefs over other people's intentions. Once we know what the other players are going to do, picking out the best choice only requires a strategic decision, a simple mathematical calculation. The problem is that the intentions of one player depend on his or her knowledge about the intentions of the other players, and these in turn depend on what they know about the intentions of the former, in a way that is characteristically circular. One of the aims of the notion of equilibrium and of its refinements is to explain how the players can restrict the range of possible intentions, on the assumption that their competitors are rational. One player can frame hypotheses on the other players' intentions grounded on the primitives of a model, this is why intentions themselves cannot be among these primitives.

In other terms, if the task of player 2 is to guess what the intended meaning is, and if she already knows which alternative is the most likely one, then there is not much to be done anymore, she only needs to multiply the subjective probability of each alternative by the payoffs that the moves available to her would yield in each of these alternatives. Suppose that p is the prior probability that player 2 assigns to the belief that player 1 wants to convey the meaning corresponding to A; and that $1-p$ is the probability of the belief that he wants to convey the meaning B. Let g_a be the gain for player 2 if she selects the interpretation A when player 1 really wants to convey A, and let m_a be her gain if she selects A when 1's intended meaning is B. Similarly, let g_b be her gain if she correctly selects B, and m_b her gain when she wrongly selects B. If we describe the situation in this way, her task is very simple, she must select A whenever $p \times g_a + (1-p) \times m_a > p \times m_b + (1-p) \times g_b$ and B whenever $p \times g_a + (1-p) \times m_a < p \times m_b + (1-p) \times g_b$. Once we know that she is able to assign a probability value to the belief that 1's intended meaning is A – no matter how she could accomplish this – there is nothing more to be explained, and hence no more need to appeal to game theory to give an account of her behaviour. But, presumably, we need game theory to explain how she could assess this probability.

This is why I claim that the content of player 1's private information has to be something more basic, and therefore that player 2's prior probabilities have to concern what player 1 actually knows. With this modelling of the game, the speaker's intention to convey a given message can be derived from facts with a minor degree of intentionality, namely his knowledge. To paraphrase Willard Van Quine (1976), it reduces the grade of *intentional involvement*. Just consider the questions 'What does player 1 know?' and 'What does player 1 want to say?'. We are not always able to provide definite answers to the questions of the first kind, but, at least, we can assess the probability of the answers, just considering what we know about the player's sources of information. Of course, we can also assess the probability of the answers to the questions of the second kind, but the data to be considered include all those relevant for the first kind, and something else, at least this person's goals. In other words, any reasoning behind an answer to a question of the first kind is conceptually simpler than that required by the second kind.

This reform imposes some restrictions on the structure of the model. If A and B are the only legitimate interpretations of an ambiguous utterance ϕ, then either he believes that A or he believes that B. But in the case we are examining, one of the two readings is a logical consequence of the other, for example we can assume that B logically entails A. If this is true, then if 1 believes that B, he necessarily believes that A. Then, as far as player 2 knows, there are two possibilities:

alternative a: 1 knows that A and it is not the case that he knows that B (either because he knows that not B, or because he does not know whether B);

alternative b: he knows that A and B.

If a is the real situation, then, if 2 selects A when 1 utters ϕ, she will acquire some new and reliable true knowledge, whose value is g_a. But, if in the same situation she chooses B, instead, she gets a false or at least unreliable new belief, and the result is m_b, and we can reasonably assume that $g_a > m_b$. If b is the real situation, then the choice of B will yield some new knowledge, let be g_b the value she puts on it. Yet, in this case, even if in she chooses A, she acquires some new knowledge. Let m_a be her payoff in this case, but since B contains more information than A, it is quite natural to assume $g_b > m_a$. We can retain Parikh's assumption that this is a game of pure coordination where, in any possible ending of the game, the payoff earned by player 1 is he same as that of player 2. We also know that, both in situation a and in situation b, player 1 can choose an unambiguous sentence to convey the same message, we can call these sentences μ_a and μ_b, respectively. In these cases, player 2 has no possibility to move. The corresponding payoffs will be g'_a and g'_b. We can assume that g_a is equal to the net value of the information carried by A when it is true and reliable minus some 'cost' c involved by the length of ϕ. g_b will have a similar composition $v_b - c$. We can conceive of cases where an unambiguous sentence is so much longer than the corresponding ambiguous one, that a cheap misunderstanding can be preferable to an unambiguous but demanding speech act. We can also imagine situations where the speakers choose ambiguous and potentially misleading messages because they do not want other people to acquire some confidential information. Just think of two spies involved in a telephone conversation, both knowing that their line has been tapped. Sometimes a leak can do more harm than a misunderstanding. I will assume that this is not the case in the conversation we are considering, and that in this case the cost of an utterance is relatively small when compared to the net value of information. This simplifying assumption entails that $g'_a > m_b$. I will also imagine that, in the conversations we are analysing, people prize transfer of information more than shunning of costs, and this entails that $g'_b > m_a$, $g'_a - m_b > g_b - g'_b$ and $g'_b - m_a > g_a - g'_a$. For the same reason, we can also safely assume that $g'_b > g'_a$.

The rationale behind the choice of a game of pure coordination where the players have the same payoff function is that when honest and rational agents communicate, they just aim at successful communication. We have to imagine that he is sincere and honest, that she believes what he says, and that this is common knowledge. For simplicity, imagine also that both of them are interested exclusively in the pure flow of information and have no further aims. This is unrealistic, of course, but it is just an idealization not more problematic than the physicist's speculations on frictionless planes. Of course there are commonly cases where this is not true, most notably when people lie. But we can legitimately focus attention on those benign cases, especially because the very possibility of lying presupposes the existence of honest communication.

Maybe the set of moves available to player 1 is incomplete. Perhaps we should also consider the possibility of uttering μ_a in situation b, and μ_b in situation a. Of course if player 1 uttered μ_b knowing that B is false, he would be lying, and, under the assumption that we are trying to analyse a case of patently honest communication, this move would yield a bad outcome for both. But the other case cannot be dismissed so easily, remember that A is true in situation b. The payoff would actually be g'_a. The fact is that whatever the choice of 2, the gain would be higher if player 1 chose μ_b, because $g'_b > g'_a$. This means that, according to the model presented here, it is never rational for player 1 to choose to utter μ_a in situation b. In technical terms, any strategy where the speaker utters μ_a in situation b or μ_b in situation a is strongly dominated, and can be eliminated from the game. The model simply predicts the existence of a scalar implicature, to the effect that if 1 utters μ_a, then 2 infers that it is not the case that 1 knows that B. Simply because the ordering among payoffs that was depicted above presupposes that if 1 knew that B, then he would not conceal this information to the hearer. In situations not covered by this analysis, the speaker could utter μ_a knowing that B, if he did not want 2 to know.

Similarly, we could include a pair of 'don't say anything' moves for player 1. Of course, when he chooses one of these, she has no possibility to move, and the payoff should be equal to 0 for both players. I will assume that both g'_a and g'_b are strictly positive. If this is the case, then, again, any strategy involving one of these additional moves is strongly dominated, hence I will ignore this possible variant of the game. Yet, this shows that I have not mentioned a fact which is implicitly presupposed by our model, namely the fact that, for example, at node a, player 1 knows that A and also *wants* to convey this information. If Parikh meant this while saying that the chance nodes 'represent [player 1's] intention to convey' A or B (Parikh, 2001, p. 27), then the objections I raised in Section 1 miss the mark. But, this does not seem to be the case. The fact that there are only two alternative states in 2's information set follows from the characteristic features of the examples considered, namely the fact that one of the two readings is entailed by the other. It is not even necessary that this be a logical entailment – like it is in our example – but the entailment has to be common knowledge. If the two alternative readings were logically and conceptually unrelated, player 2 would have an information set containing three elements. And of course we can conceive of cases where an ambiguous sentence admits of more than two readings. In (2006, p. 351) Parikh analyses a case of disambiguation where the alternative readings are unrelated, building a model with only two chance events. According to the line pursued here, these cases require models with three chance events.

We can retain the diagram of Figure 1 as the extensive form of a disambiguation game for sentences like (1) and (2), we only need to change the interpretation of the chance nodes and of 1's private information.

Still, one could ask: Why a game? In other words, why should not 2's choice be just a strategic decision drawn upon her prior probabilities of the chance events a and b? First of all, the decision-theoretic problem that has been depicted manifests the characteristic circularity that imposes a game-theoretic analysis. What is the best choice for 1, when he is in a? It is I, if 2 will choose A, and E otherwise. So the question is, what shall *she* do? Or, equivalently, what is the best choice for *her*? When it is up to her to make a move, she knows that 1 has chosen to use the ambiguous sentence ϕ. The best choice for her is A, if 1 has planned to choose I when in a, and e when in b, for example. So, his best choice depends on her best choice, and vice versa. Second, even if the plan to act strategically upon the prior probabilities p and $1-p$ is not outright irrational on the part of 2, after all it is an equilibrium strategy, it is not the best our two players can achieve from the point of view of efficiency. Since there is no guarantee that the corresponding equilibrium is the most efficient one.

3 Equilibria

I will now establish a few properties of the model.

Theorem 3.1 *Ii is strongly dominated in G.*

Proof. Ii is strongly dominated if and only if $\exists \sigma_1 \in \Delta(C_1)$ such that

$$\begin{aligned} &u(Ii, A) < \sigma_1(Ei)u(Ei, A)+ \\ &\sigma_1(Ee)u(Ee, A) + \sigma_1(Ie)u(Ie, A)+ \\ &(1 - \sigma_1(Ei) - \sigma_1(Ee) - \sigma_1(Ie))u(Ii, A) \end{aligned} \tag{3}$$

and

$$\begin{aligned} &u(Ii, B) < \sigma_1(Ei)u(Ei, B)+ \\ &\sigma_1(Ee)u(Ee, B) + \sigma_1(Ie)u(Ie, B)+ \\ &(1 - \sigma_1(Ei) - \sigma_1(Ee) - \sigma_1(Ie))u(Ii, B) \end{aligned} \tag{4}$$

Inequalities (3) and (4) are equivalent to

$$\frac{\sigma_1(Ei)+\sigma_1(Ee)}{\sigma_1(Ie)+\sigma_1(Ee)} < \frac{(1-p)(g'_b - m_a)}{p(g_a - g'_a)} \tag{5}$$

and

$$\frac{\sigma_1(Ei)+\sigma_1(Ee)}{\sigma_1(Ie)+\sigma_1(Ee)} > \frac{(1-p)(g_b - g'_b)}{p(g'_a - m_b)} \tag{6}$$

respectively. Given the ordering among payoffs stated in Section 2,

$$\frac{g'_b - m_a}{g_a - g'_a} > 1 > \frac{g_b - g'_b}{g'_a - m_b}$$

and hence

$$\frac{(1-p)(g'_b - m_a)}{p(g_a - g'_a)} > \frac{(1-p)(g_b - g'_b)}{p(g'_a - m_b)}$$

At this point it is an easy task to find values for $\sigma_1(Ei)$, $\sigma_1(Ie)$, and $\sigma_1(Ee)$ that satisfy inequalities (5) and (6). QED

Observe that Theorem 3.1 entails that no strategy profile τ where $\tau_1(Ii) > 0$ is a Nash equilibrium.

Theorem 3.2 *There is no equilibrium in G where both Ei and Ie have strictly positive probability.*

Proof. Assume that σ is such an equilibrium. Then the following inequalities have to be true:

$$\begin{aligned}
\sum_{c_2 \in C_2} \sigma_2(c_2) u(Ei, c_2) &\geq \sum_{c_2 \in C_2} \sigma_2(c_2) u(Ee, c_2) \\
\sum_{c_2 \in C_2} \sigma_2(c_2) u(Ie, c_2) &\geq \sum_{c_2 \in C_2} \sigma_2(c_2) u(Ee, c_2)
\end{aligned}$$

They are equivalent to

$$\begin{aligned}
\sigma_2(A) &\leq \frac{g_b - g'_b}{g_b - m_a} \\
\sigma_2(A) &\geq \frac{g'_a - m_b}{g_a - m_b}
\end{aligned}$$

respectively. But this cannot be. In fact, given the ordering among payoffs of Section 2,

$$\begin{aligned}(g'_a - m_b)(g'_b - m_a) &> (g_b - g'_b)(g_a - g'_a),\\ (g'_a - m_b)(g_b - g'_b) + (g'_a - m_b)(g'_b - m_a) &> \\ > (g_b - g'_b)(g_a - g'_a) &+ (g'_a - m_b)(g_b - g'_b),\\ (g'_a - m_b)(g_b - m_a) &> (g_b - g'_b)(g_a - m_b)\end{aligned}$$

Hence

$$\frac{g'_a - m_b}{g_a - m_b} > \frac{g_b - g'_b}{g_b - m_a} \tag{7}$$

QED

Theorem 3.3 *There is no equilibrium G where both Ie and Ee have strictly positive probability.*

Proof. Assume that σ is such an equilibrium. Then the following equation has to be true

$$\sum_{c_2 \in C_2} \sigma_2(c_2) u(Ie, c_2) = \sum_{c_2 \in C_2} \sigma_2(c_2) u(Ee, c_2)$$

which amounts to

$$\sigma_2(A) = \frac{g'_a - m_b}{g_a - m_b}$$

This means that $1 > \sigma_2(A) > 0$, hence in this equilibrium player 2 is indifferent between strategies A and B, and this means

$$\sum_{c_1 \in C_1} \sigma_1(c_1) u(c_1, A) = \sum_{c_1 \in C_1} \sigma_1(c_1) u(c_1, B) \tag{8}$$

Since $\sigma_1(Ei) = 0$ and $\sigma_1(Ii) = 0$ because of Theorems 3.1 and 3.2, (8) becomes $g_a = m_b$, which is impossible. QED

Theorem 3.4 *There is no equilibrium where both Ei and Ee have strictly positive probability.*

Proof. Analogous to the preceding one. QED

How many equilibria are there? Of course there are two equilibria in pure strategies, namely $\eta = ([Ie], [A])$ and $\theta = ([Ei], [B])$, but there is also an infinite set of mixed equilibria.

Theorem 3.5 *If*

$$\pi_1(Ee) = 1 \tag{9}$$

and

$$\frac{g'_a - m_b}{g_a - m_b} \geq \pi_2(A) \geq \frac{g_b - g'_b}{g_b - m_a} \tag{10}$$

then π is a Nash equilibrium.

Proof. Consider a modified game $G^* = \{N, C_1^*, C_2, u^*\}$ where

$$C_1^* = \{Ei, Ie, Ee\}$$

and u^* is just u after its domain has been restricted accordingly. Since Ii is strongly dominated because of Theorem 3.1, every equilibrium of G^* is an equilibrium of G, and vice versa. Suppose that π is a strategy profile that satisfies conditions (9) and (10). Define ω as $p(g'_a - g'_b) + g'_b$, which is the expected payoff of both players under π. Since player 2 is clearly indifferent between A and B when player 1's strategy is $[Ee]$, in order to show that π is an equilibrium, we only need to prove the following statements:

$$\omega \geq \sum_{c_2 \in C_2} \pi_2(c_2) u(Ei, c_2) \tag{11}$$

$$\omega \geq \sum_{c_2 \in C_2} \pi_2(c_2) u(Ie, c_2) \tag{12}$$

But the conjunction of conditions (11) and (12) is equivalent to (10). Hence π is a Nash equilibrium of G^* and therefore of G as well. QED

Theorems 3.1, 3.2, 3.3, and 3.4 entail that there are no other equilibria. Are they trembling hand perfect? All the strategies $[Ei]$, $[Ie]$, $[A]$, and $[B]$ are visibly undominated, and this entails that both pure equilibria are perfect (Osborne and Rubinstein, 1994, prop. 248.2). $[Ee]$ is not weakly dominated either, hence the mixed equilibria are perfect as well, but this might not be perceived at first sight.

Theorem 3.6 $[Ee]$ *is not weakly dominated.*

Proof. If $[Ee]$ is weakly dominated, then, for some $\sigma_1 \in \Delta(C_1)$,

$$\forall \pi_2 \in \Delta(C_2), u([Ee], \pi_2) \leq u(\sigma_1, \pi_2)$$

i.e.

$$\begin{aligned} &\forall \pi_2 \in \Delta(C_2), \\ &u([Ee], \pi_2) \leq \sum_{c_1 \in C_1} \sigma_1(c_1)[\pi_2(A)u(c_1, A) + (1 - \pi_2(A))u(c_1, B))] \end{aligned}$$

If we instantiate with

$$\pi_2(A) = \frac{g'_a - m_b}{g_a - m_b}$$

this becomes

$$\frac{g'_a - m_b}{g_a - m_b} \leq \frac{g_b - g'_b}{g_b - m_a}$$

which contradicts (7) QED

4 Perfect Equilibria in Extensive Form

One might hope to select a unique equilibrium arguing that in our analysis player 2 does not exploit all the evidence she has at her disposal, since in order to make a rational choice she must consider not the prior probability of a and b, but the conditional probability of those events, given that player 1 decided to utter ϕ. This suggests that we search for perfect equilibria in the extensive form of the game. In this section I show that this is of no help, because all of the Nash equilibria of the normal representation correspond to perfect equilibria of the extensive form.

The multiagent representation (Myerson, 1991) – also called agent-normal form (Selten, 1975), and agent strategic (Osborne and Rubinstein, 1994) – is a way of representing games in extensive form as games in strategic form, alternative to the normal representation. In the multiagent representation of some extensive-form game Γ^e, there is a player, called (temporary) agent, for every information set of every player of Γ^e. Hence, as far as our game is concerned, player 1 is represented by two agents in the multiagent representation, say a and b. While there is only one agent for player 2, say c.

	e	
	A	B
E	$p \times g'_a + (1-p) \times g'_b$	$p \times g'_a + (1-p) \times g'_b$
I	$p \times g_a + (1-p) \times g'_b$	$p \times m_b + (1-p) \times g'_b$

	i	
	A	B
E	$p \times g'_a + (1-p) \times m_a$	$p \times g'_a + (1-p) \times g_b$
I	$p \times g_a + (1-p) \times m_a$	$p \times m_b + (1-p) \times g_b$

Table 2: Disambiguation game: the multiagent representation

The multiagent representation of our disambiguation game is represented in Table 2.

A *behavioural strategy profile* of a game in extensive form is a mixed strategy profile of its multiagent representation. Let 'G^e' be the name of the extensive form of the disambiguation game. A generic behavioural strategy profile of G^e is $(\sigma_a, \sigma_b, \sigma_c)$, and it specifies a probability distribution for every agent of every player. The behavioural strategy profile $([I], [e], [A])$ corresponds to our Nash equilibrium η in an intuitive way, so that it can be called its *behavioural representation* (Myerson, 1991). Since there should not be any danger of misunderstanding, until the end of this section, I will use the names of the strategy profiles of (the normal representation) G to refer to their behavioural representations in G^e. Hence, I will set $\eta = (\eta_a, \eta_b, \eta_c) = ([I], [e], [A])$, and similarly for the other equilibria.

Definition 4.1 A trembling hand perfect equilibrium of a game in extensive form is a trembling hand perfect equilibrium of its multiagent representation (Myerson, 1991; Osborne and Rubinstein, 1994). $\lhd$

Theorem 4.2 *η is a trembling hand perfect equilibrium of G^e*

Proof. Recall that η is a perfect equilibrium iff there exists a sequence $(\eta^k)_{k=1}^{\infty}$ such that each η^k is a perturbed behavioural strategy profile where every move gets positive probability, and, moreover

(i)

$$\lim_{k \to \infty} \eta_s^k(d_s) = \eta_s(d_s) \quad \forall s \in S \quad \forall d_s \in D_s$$

(ii)

$$\begin{aligned}&\eta_s \in \operatorname{argmax}_{\tau_s \in \Delta(D_s)}\\&\sum_{d\in D}\left(\prod_{r\in N-s}\eta_r^k(d_r)\right)\tau_s(d_s)u(d)\\&\forall s \in S\end{aligned}$$

where $S = (a, b, c)$ is the set of all information states of all players, and, for each $s \in S$, D_s is the set of moves available to the relevant player in state s, and $D = \times_{s\in S} D_s$. It is not difficult to find a sequence satisfying these criteria. Set

$$\xi = \frac{(1-p)(g_b - m_a)}{p(g_a - m_b)}$$

Then $\forall k \in (1, 2, 3, ...)$, if $\xi \geq 1$,

$$\eta_a^k(I) = \frac{2k-1}{2k} \quad \eta_b^k(i) = \frac{1}{2k\xi} \quad \eta_c^k(A) = 1 - \frac{g_a - g'_a}{k(g_a - m_b)}$$

If $\xi < 1$, instead, set

$$\eta_b^k(i) = \frac{1}{2k}$$

and the rest as before. You can see at a glance that these sequences satisfy condition (i). Consider now the expected payoff of player 1 when he is in state a and is planning to make move $\tau_a \in \Delta(D_a)$, and all other agents behave according to scenario η^k. It is equal to

$$\begin{aligned}&\sum_{d-a\in D-a}\left(\prod_{r\in N-a}\eta_r^k(d_r)\right)\times\\&[\tau_a(I)u(d_{-a}, I) + (1-\tau_a(I))u(d_{-a}, E)]\end{aligned} \tag{13}$$

We can consider (13) as a function of $\tau_a(I)$, and if we calculate the derivative of this function we get

$$p[\eta_c^k(A)(g_a - m_b) + m_b - g'_a]$$

As you can easily verify, this value is either null or positive for all k, and this means that, since $\eta_a(I) = 1$

$$\eta_a \in \text{argmax}_{\tau_a \in \Delta(D_a)}$$
$$\sum_{d \in D} \left(\prod_{r \in N-a} \eta_r^k(d_r) \right) \tau_a(d_a) u(d)$$

Similarly, if you consider the corresponding expected payoff for player 1 when he is in state b, i.e.

$$\sum_{d-b \in D-b} \left(\prod_{r \in N-b} \eta_r^k(d_r) \right) \times$$
$$[\tau_b(i) u(d_{-b}, i) + (1 - \tau_b(i)) u(d_{-b}, e)]$$

regard it as a function of $\tau_b(i)$, and calculate its derivative, you get

$$(1-p)[\eta_c^k(A)(m_a - g_b) + g_b - g_b']$$

which is either null or negative for all k, because of inequality (7), and this means that, since $\eta_b(i) = 0$,

$$\eta_b \in \text{argmax}_{\tau_b \in \Delta(D_b)}$$
$$\sum_{d \in D} \left(\prod_{r \in N-b} \eta_r^k(d_r) \right) \tau_b(d_b) u(d)$$

Finally, if you calculate the expected payoff for player 2, you have

$$\sum_{d-c \in D-c} \left(\prod_{r \in N-c} \eta_r^k(d_r) \right) \times$$
$$[\tau_c(A) u(d_{-c}, A) + (1 - \tau_c(A)) u(d_{-c}, B)]$$

whose derivative is

$$\eta_a^k(I) p(g_a - m_b) + \eta_b^k(i)(1-p)(m_a - g_b)$$

which is either null or positive for all k, and this entails

$$\eta_c \in \text{argmax}_{\tau_c \in \Delta(D_c)}$$
$$\sum_{d \in D} \left(\prod_{r \in N-c} \eta_r^k(d_r) \right) \tau_c(d_c) u(d)$$

QED

The case of θ is completely analogous.

Theorem 4.3 *θ is a trembling hand perfect equilibrium of G^e*

Proof. A suitable sequence is

$$\theta_a^k(I) = \frac{1}{2k} \quad \theta_b^k(i) = \frac{2k-1}{2k} \quad \theta_c^k(A) = \frac{g_b - g_b'}{k(g_b - m_a)}$$

if $\xi \geq 1$, and

$$\theta_a^k(I) = \frac{\xi}{2k} \quad \theta_b^k(i) = \frac{2k-1}{2k} \quad \theta_c^k(A) = \frac{g_b - g_b'}{k(g_b - m_a)}$$

if $\xi < 1$. QED

As for the mixed equilibria the case is simpler.

Theorem 4.4 *The mixed equilibria π are trembling hand perfect in the extensive form of the game*

Proof. Since $1 > \pi_c(A) > 0$, we can set $\pi_c^k(A) = \pi_c(A)$, and

$$\eta_a^k(I) = \frac{1}{2k} \quad \eta_b^k(i) = \frac{1}{2k\xi}$$

whenever $\xi \geq 1$, and

$$\eta_a^k(I) = \frac{\xi}{2k} \quad \eta_b^k(i) = \frac{1}{2k}$$

otherwise. QED

5 Efficiency

Summing up, in the strategic form of the game, there are two equilibria in pure strategies, namely η and θ, and many mixed equilibria π, and all are trembling hand perfect. All the mixed equilibria are somehow equivalent, since they yield the same expected payoff, and they all amount to the fact that player 1 goes for the costly but unambiguous option, and player 2 has no opportunity to move. These mixed equilibria are the least efficient ones.

As for the equilibria in pure strategies, η is the unique Pareto efficient equilibrium iff

$$p > \frac{g_b - g'_b}{g_b - g'_b + g_a - g'_a} \tag{14}$$

and θ is the unique Pareto efficient equilibrium iff

$$p < \frac{g_b - g'_b}{g_b - g'_b + g_a - g'_a} \tag{15}$$

If we assume that μ_a and μ_b have analogous length, then the actual critical value

$$\frac{g_b - g'_b}{g_b - g'_b + g_a - g'_a}$$

will be approximately equal to 1/2. Parikh's account predicts that the players will tend to converge on the most efficient equilibrium, and I will accept this view, which seems to be empirically adequate, at least at first sight. But Robert Van Rooij rejects this solution concept, claiming that it is 'unusual' (2004, p. 506). This claim is quite odd. First, there is some agreement among some scholars on the view that we should expect rational players to converge on efficient equilibria in many kinds of games (Harsanyi and Selten, 1988; Myerson, 1991). Second, we should subscribe this remark of Robert Aumann (2000, p. 5).

> My main thesis is that a solution concept should be judged more by what it does than by what it is; more by its success in establishing relationships and providing insights into the workings of the social processes to which it is applied than by considerations of *a priori* plausibility based on its definition alone.

Yet, the doctrine is incomplete, as it does not explain what should happen in the limit case where

$$p = \frac{g_b - g'_b}{g_b - g'_b + g_a - g'_a} \tag{16}$$

that makes both η and θ (weakly) Pareto efficient. I will now provide an answer to this question. We can reasonably expect that, in these cases, 1 will choose strategy $[Ee]$, inefficient but safe, and since 2 can anticipate this, she will be indifferent between her two options A and B. Even in this case, I will just assume that this claim is empirically adequate. I will first put forward a philosophical justification of the doctrine, and then provide a formal definition.

I will start with a philosophical parable. Let us admit that the speakers will converge on the unique Pareto dominant equilirium, whenever there is one. This kind of coordination is very profitable for both, we could suspect that this is not possible without a previous agreement, a kind of *social contract*. Pursuing this fantasy a little further, we can even speculate over the behaviour the speakers had before they entered such an imaginary contract. We can fairly assume that it would have been in line with an equilibrium, a strategy not in equilibrium is not rational. The only equilibria when there is no coordination are the mixed ones, costly but safe. Given the opportunity to make an agreement, they would have decided to converge on the most efficient equilibrium, whenever possible. If we agree on the upshot of this hypothetical bargain, it does not have to be real in order to have visible effects. Both players are able to deduce what would have happened in such a counterfactual situation, because this can be inferred from the structure of the game, it is a feature of the game, which is common knowledge.

This kind of argument hinges on what is called 'preplay communication' (Myerson, 1991, pp. 109-113), and, according to Parikh it is untenable, for two reasons. First, if you explain successful communication in terms of preplay communication you fall into an infinite regress. Second, 'even if such an infinite regress were avoidable, the solution would certainly require a great deal of effort suggesting that languages aren't quite so efficient as they in fact are' (Parikh, 2001, p. 39n). I argue that both of these tenets can be rejected. The model presented here is an account of disambiguation, which is a particular phenomenon occurring in communication. I claimed that our two players could converge on a unique equilibrium, if they considered what would have happened if they had had the opportunity to reach an agreement over a coordinated plan. If this imaginary preplay communication is conceived as involving only unambiguous sentences, there seems to be no danger of an infinite regress, yet the response is the same: they would have agreed to converge on the unique Pareto efficient equilibrium. The second point is less clear to me, since the kind of reasoning that we attribute to our players does not seem to involve a great deal of computational effort, compared to the construction of the model itself.

Let us now go back to the social contract fantasy, to the point where speaker and hearer play a disambiguation game *before* the making of the contract. Let us imagine that $([Ie],[A])$ is strongly Pareto dominant in this game. Assume that 2 predicts that 1 will opt for $[Ee]$. Given this belief, she is indifferent between A and B, yet she at least has a reason to choose A instead of B. She can think: 'If my expectation concerning 1's behaviour is actually wrong, i.e. if he is going to deviate toward one of the strategies of the pure equilibria, quite likely it will be $[Ie]$, because it is dominant, not $[Ei]$.' But 1 himself can suspect that 2 will act upon this reasoning, and this would really lead him to deviate toward the dominant equilibrium. This kind of attraction on the part of dominant equilibria is an instance of the so-called *focal-point effect* (Myerson, 1991, pp. 108-114). But if in this kind of game this effect is triggered by the reasoning outlined above, the effect will actually occur only when there is a unique Pareto dominant equilibrium, otherwise the players will be stuck in the imaginary primeval condition antedating the social contract.

Here ends the parable and begins the formal definition. My speculation hinged on the action undertaken by 2, if 1, unexpectedly, deviates from his equilibrium strategy $[Ee]$. Her actions depend on her beliefs, therefore the question becomes: How shall she revise her beliefs if 1 deviates from his equilibrium strategy? This is tantamount to the question: What is the probability of a conditioned on the evidence that 1 chose I or i? The problem is of course that this conditional probability is left undefined by traditional Bayesian probability theory, because the condition has null prior probability, under this equilibrium. This kind of question is actually central in past and current debates in game theory, since it motivates most equilibrium refinements, but we have seen that these are of no help here.

We should divert attention toward the class of *signaling games*. In a signaling game we have two players, a *sender* and a *receiver*. The sender knows his *type* which is drawn from a set of possible types, according to some prior probability which is common knowledge. The sender must send a message to the receiver. The receiver does not observe the sender's type, but she sees his message. Finally, the receiver must chose her action, and the game ends. The payoff of both players depend on the sender's type, the message chosen by the sender, and the action performed by the receiver. In our disambiguation game, player 1 is the sender, a and b are his possible types, his messages are his moves – E and I or e and i, depending on type – A and B are the actions available to 2.

None of the most common solution concepts that have been developed for signaling games can rule out the undesired equilibria in our disambiguation game, all of them leave the situation unchanged. I will therefore sketch a new solution concept, that subsumes various existing solutions, most notably the 'Intuitive Criterion' (Cho and Kreps, 1987), and 'Divinity' (Banks and Sobel, 1987), and provides a non-arbitrary way of calculating conditional probabilities in off-the-path information states. This solution concept is new, to my knowledge, yet it was inspired by the papers on signaling games cited in this section.

If 1 is in situation a, his equilibrium strategy under a mixed equilibrium π will give him a payoff equal to g'_a. Let $D(a)$ be the set of mixed strategies of player 2 that would make a deviation from the equilibrium at least as good as his equilibrium strategy, i.e.

$$D(a) = \{\varphi_2 : g'_a \leq \varphi_2(A)g_a + (1 - \varphi_2(A))m_b\}$$

Analogously

$$D(b) = \{\varphi_2 : g'_b \leq \varphi_2(A)m_a + (1 - \varphi_2(A))g_b\}$$

Clearly

$$D(a) = \left\{\varphi_2 : \frac{g'_a - m_b}{g_a - m_b} \leq \varphi_2(A)\right\}$$

$$D(b) = \left\{\varphi_2 : \frac{g_b - g'_b}{g_b - m_a} \geq \varphi_2(A)\right\}$$

Consider now the Lebesgue measures of these two sets, namely

$$\lambda(D(a)) = 1 - \frac{g'_a - m_b}{g_a - m_b} = \frac{g_a - g'_a}{g_a - m_b}$$

and

$$\lambda(D(b)) = \frac{g_b - g'_b}{g_b - m_a}$$

I claim that $\lambda(D(a))$ is relevant to the probability that 1 will choose move I when he is in a, in the following way. Speaking figuratively, my hypothesis is that the value $\lambda(D(a))$ is proportional to the infinitesimal probability that 1 will deviate from his equilibrium strategy, when he is in situation a. Similarly, that $\lambda(D(b))$ is proportional to the infinitesimal probability that 1 will deviate from his equilibrium strategy, when he is in situation b. We know that, under π, the probability that 1 will choose I or i is equal to 0. Hence, making this metaphor more formal, I imagine that the probability that the path of the game will go through one of the nodes in 2's information set, conditioned on the event that 1 is in a, is, under the equilibrium π, equal to $\varepsilon\lambda(D(a)$, where ε is some infinitesimal coefficient. Similarly for situation b. None of this has any effect on on-the-path probabilities, in fact $lim_{\varepsilon\to 0}\varepsilon\lambda(D(a)) = 0$. But it gives a non-arbitrary way of defining off-the-path conditional probabilities, since we can exploit a form of Bayes' Theorem to define the conditional probability of the event that 1 is in a, on the condition that he has chosen I or i. More formally, let ι be the event that 1 has chosen to use the ambiguous sentence ϕ. When x and y are two events, let us use the expression '$P_x(y)$' to denote the probability of y conditioned on x. If

$$P_a(\iota) = \varepsilon\lambda(D(a))$$

$$P_b(\iota) = \varepsilon\lambda(D(b))$$

then, since p and $1-p$ are the prior probabilities of events a and b,

$$P_\iota(a) = \frac{p\varepsilon\lambda(D(a))}{p\varepsilon\lambda(D(a)) + (1-p)\varepsilon\lambda(D(b))}$$

$$P_\iota(a) = \frac{p\lambda(D(a))}{p\lambda(D(a)) + (1-p)\lambda(D(b))}$$

Similarly,

$$P_\iota(b) = \frac{(1-p)\lambda(D(b))}{p\lambda(D(a)) + (1-p)\lambda(D(b))}$$

The reader can check that, given these beliefs, player 2 strictly prefers A over B iff (14) holds, she srtictly prefers B iff (15) holds, and she is indifferent iff (16) holds. This entails that the criterion newly introduced breaks all the mixed equilibria whenever one of the pure ones is strictly Pareto dominant, and leaves them intact otherwise. The idea of a link between Lebesgue measures of sets like $D(a)$ and $D(b)$ in signaling games and the probability of a deviation from equilibrium is due to Gonzalo Olcina (1997), only the hypothesis that they are to be proportional to these probabilities, and that they should be used to calculate off-the-path conditional probabilities is new.

Summing up, this solution suggests that players will always try to conform to a strategy of some mixed equilibrium in a disambiguation game, and that they will fail whenever this equilibrium does not satisfy the above criterion, i.e. almost always, in which case they will end up converging toward the unique dominant equilibrium.

In order for this new solution concept to be viable, it has to be proved that given any game, or any game belonging to some suitable class, this solution will always pick out a non-empty set of plausible equilibria, for example that it will always select at least one sequential equilibrium. I leave this as an open question. Let me just observe that this solution is not patently ad hoc, since it can be applied to a class of games much wider than that of disambiguation games.

6 Conclusion

Summing up, the substance of this work is a new game-theoretic analysis of the capacity humans have to communicate using ambiguous expressions. The background hypothesis is that these tasks are accomplished because humans are rational creatures, and, when two people are involved in a conversation, they crucially capitalize on this fact, assuming that it is common knowledge. I built on ideas first developed by Prashant Parikh, raising some objections that led me to modify his models.

I built a game of imperfect information in extensive form, where a hearer and a speaker are the two players, the speaker has some private information, and his task is to convey this piece of information to the hearer. Here lies the main difference between my analysis and Parikh's, since, in his model, the relevant private information of the speaker is the intended meaning of his speech act, while in mine the private information is just some piece of knowledge that he wants to share with the hearer. I argued that my reform renders the theory more natural and conceptually simpler.

The examples I chose as sample cases were simpler to analyse than more general cases, because of the structural features of the resulting model. In the end I retain Parikh's conclusion that speakers tend to focus on efficient equilibria, but I also proposed a solution to a problem that had been left open, namely, the strategy adopted by the speakers when there is not a unique efficient equilibrium. I argued that, in this case, the speaker goes for the ambiguous expression, which is costly, but safe. The intuitive argument I used to back both of these tenets hinges on the idea that the players are able to guess the joint strategy they would agree on, were they allowed some preplay communication before the beginning of the game. This kind of argument is not new. It is crucial that the players do not really need to entertain this kind of communication in order to know what would ensue from it. The formalisation of this intuitive argument hinges on a solution concept that defines conditional probabilities in information states that have null prior probability. This solution concept can be applied to a variety of games much wider than the limited set of disambiguation games, most notably classic signaling games like 'Beer or Quiche'.

Now there are two directions where this research can be pushed forward. First, one has to apply the methods and ideas presented here to a wider class of conversational games, starting from other disambiguation games. Second, the range of games where the new solution concept proposed here can be applied has to be clearly delineated.[1]

References

Aumann, R. J. (2000). *Collected Papers.* MIT Press, Cambridge Massachusetts.

Banks, J. S. and Sobel, J. (1987). Equilibrium selection in signaling games. *Econometrica*, **55**(3), 647–661.

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

Harsanyi, J. and Selten, R. (1988). *A General Theory of Equilibrium Selection in Games.* MIT Press, Cambridge Massachusetts.

Myerson, R. (1991). *Game Theory: Analysis of Conflict.* Harvard University Press, Cambridge Massachusetts.

Olcina, G. (1997). Forward induction in games with an outside option. *Theory and Decision*, **42**(2), 177–192.

Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory.* MIT Press, Cambridge Massachusetts.

[1] I wish to thank several people to whom early versions of this work have been presented, during lectures delivered at the University of Milan and at the University of Siena in May and June 2007.

Parikh, P. (1992). A game-theoretic account of implicature. In Y. Moses, editor, *Theoretical Aspects of Reasoning about Knowledge*.

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

Parikh, P. (2006). Radical semantics: A new theory of meaning. *Journal of Philosophical Logic*, **35**(4), 349–391.

Quine, W. V. (1976). *The Ways of Paradox and other Essays*. Harvard University Press, Cambridge Massachusetts, 2nd edition.

Selten, R. (1975). Reexamination of the perfectness concepts for equilibrium points in extensive games. *International Journal of Game Theory*, **4**(1), 25–55.

Spence, M. (1975). Job market signaling. *The Quarterly Journal of Economics*, **87**(3), 355–374.

Van Rooij, R. (2004). Signalling games select Horn strategies. *Linguistics and Philosophy*, **27**(4), 493–527.