

---

# A Flexible Mechanism for Dialogue Design

---

**Guido Boella**

Dipartimento di Informatica  
Università di Torino  
Italia  
guido@di.unito.it

**Jelle Gerbrandy**

Dipartimento di Informatica  
Università di Torino  
Italia  
jelle@gerbrandy.com

**Joris Hulstijn**

Faculty of Economics  
and Business Administration  
Vrije Universiteit Amsterdam  
Nederland  
jhulstijn@feweb.vu.nl

## Abstract

In this paper we apply insights from mechanism design to the design of agent interaction protocols. We show how this allows us a more flexible approach to the design of agent interaction protocols. By way of an analysis of a dialogue game from the literature we show how a protocol with many constraints on the moves allowed can be replaced by one in which we relax the rules, but make stronger assumptions on the type of participant that is involved in the dialogue. We can then use techniques from game theory and mechanism design to show that many of the constraints in the original protocol can be derived as properties of rational behavior.

## 1 Introduction

Mechanism design is the art of designing a protocol to achieve some desirable outcome. For example, a protocol must be shown to terminate, or produce a fair outcome for all participants. This is achieved by setting up a structure in which each participant has an incentive to behave in such a way that the desired outcome is reached. If this is so, the protocol is said to implement the desired outcome.

The protocol must implement the desired outcome for all potential participants, and usually few assumptions are made about these participants. That means that the resulting mechanism must place relatively strong constraints on the behavior of players – that is, limit the moves allowed – to force a wide range of types of players to behave in the intended way.

In some cases, however, we may not be able or want to enforce strict protocol rules, but we are in a position to make strong assumptions on the type of participants in the protocol. To give a very simple example: suppose

the game is tic-tac-toe, and we want to implement a mechanism such that the game will always end in a draw. One way to enforce a draw for all player types is to simply forbid all moves that let either player win. But if we make certain assumptions on the players – that they want to win, that they have enough reasoning capabilities to reason about all possible outcomes – we do not need to impose any extra rules: with such players, we already know that they will play in such a way that the result is a draw.

This situation – less rules, more knowledge about player types – is of particular interest in interaction and communication protocols. In an open environment participants can not be assumed to share a complete and fixed interaction protocol. So researchers are looking for more open or flexible approaches to agent interaction (Yolum and Singh, 2002; Mazouzi et al., 2002). For each interaction purpose, agents may for example download partial protocols from a repository, and jointly construct more complex protocols to suit their needs. Alternatively, designers may develop context dependent protocols, to be selected at runtime. In such scenarios, an important parameter to select appropriate (components of) protocols will be the *dialogue type*, such as negotiation, information seeking, or debate (Walton and Krabbe, 1995, p 65).

We can use the fact that participants want to engage in an interaction in a specific *role* to make assumptions about their expectations and preferences about the outcome of the interaction. We show in this paper how we can use such assumptions to predict the behavior of the participants in one particular example.

To illustrate this idea, we take a relatively restricted protocol for *persuasion dialogue* from the literature (Mackenzie, 1979), and show how many of the dialogue rules need not be explicitly imposed, but can be derived as properties of rational behaviour, given certain assumptions about the preferences of the players based on their roles. The purpose of the paper is to demonstrate the feasibility of this approach.

In section 2 we review some basic definitions from mechanism design. In section 3 we define and discuss our mechanism for persuasion dialogue, and show how it implements many of the rules from Mackenzie (1979). The paper ends with a comparison with similar research, and some ideas for future work. Details about the proofs can be found in the appendix.

## 2 Basic definitions

We start with a quick review of a number of standard definitions from game theory and mechanism design, adapting terminology to suit our purposes.

A *protocol* (or *mechanism*) is essentially a game frame, in which a number of *participants* can choose, alternately, one of a set of available *moves* until a certain *outcome* is reached. Specific participants will have *preferences* over these outcomes, and a protocol together with a specification of these preferences defines a *game*.

### Definition 1 (Protocols)

A *protocol*  $P$  is a tuple  $(S, \longrightarrow, s_0, \text{roles}, \tau)$  where  $(S, \longrightarrow, s_0)$  is a tree with  $s_0$  as its root,  $\text{roles}$  a set of *roles*, and  $\tau$  is a turn-taking function which assigns to each non-final state  $s$  an element from  $\text{roles}$ . Intuitively, it is the player of the role  $\tau(s)$  that is to choose among the branches leaving  $s$ .

An *outcome* in  $P$  is a maximal path in  $P$ .

A *preference order* (or *type*) for a protocol  $P$  is total ordering on its outcomes. We write  $s \preceq_r s'$  if the player in role  $r$  prefers  $s$  over  $s'$ .

A *role specification* (or *type space*)  $\Theta = (\theta_r)_{r \in \text{roles}}$  is a set of preference orderings  $\theta_r$  for each role  $r$ .  $\square$

A protocol together with a role specification can be seen as a *game of imperfect information*. We assume that the players know their own preferences, but know no more about that the preferences of the other players as is given by the role specification. And we assume that all of this is common knowledge. We diverge somewhat from standard practice in game theory by identifying outcomes with paths in the game. In mechanism design it is useful to identify a set of outcomes that are independent of the structure of the game itself. As we are interested in communication games, we are also interested in the way the outcome is reached, and we have taken a shortcut by defining outcomes by plays in the game itself. In terms of the persuasion dialogue that we will introduce later: we are interested in the speech acts that are made during the dialogue just as much as in the final commitments of the players.

We can use tools from *game theory* to predict how rational players – with specific preferences – will behave within a protocol. Even if game theory does not give us a general answer to the question what a rational player will do in any kind of strategic situation, it does give us a number of *solution concepts*: ways of defining what constitutes rational behavior in a given situation.

Without claiming that this is the most appropriate notion, in this paper, we will use what is perhaps the most famous among these solution concepts: the notion of a Nash Equilibrium. A suitable alternative solution concept would be survival after iterated elimination of dominated strategies. Like dialogues, this solution concept is constructive: it partly depends on how the outcome is reached. For players with bounded rationality, only outcomes that can actually be constructed are relevant. Despite this apparent suitability, we have selected the Nash equilibria solution concept for this paper, because it is well known, and relatively easy to make illustrative proofs.

**Definition 2 (Game Theory)** Let  $G$  be a protocol,  $\Theta$  a role specification.

A *strategy*  $\sigma_r$  for role  $r$  is a function that assigns to each  $\theta_r \in \Theta_r$  and each state  $s$  in  $G$  such that  $\tau(s) = r$ , a move in  $s$  (that is, a successor state of  $s$ ).

A *strategy profile*  $\sigma$  is a tuple of strategies, one for each role. Given a specific preference profile  $\theta \in \times_{r \in \text{roles}} \Theta_r$ , a strategy profile determines a unique maximal path in the game, and we write  $\text{outcome}(\sigma(\theta))$  for this value.

Moreover, since  $\sigma$  defines a choice in each non-terminal node, it identifies a unique outcome given any state in the tree: we write  $\text{outcome}(\sigma, s)$  for that path.

We say that a strategy  $\sigma_r$  is a *best response* in a strategy profile  $\sigma$  iff for all  $\theta \in \Theta$ , and for all strategies  $\tau_r$  for  $r$ , the outcome of  $\sigma$  given  $\theta$  is at least as preferred by  $r$  as the outcome, given  $\theta$ , of the strategy profile  $\sigma[r/\tau_r]$  obtained by replacing  $\sigma_r$  in  $\sigma$  by  $\tau_r$ .

A strategy profile is a *Nash equilibrium* in a game  $G$  just in case each strategy in the profile is a best response.

We write  $\tau[r/\sigma_r]$  for the strategy profile that results if we replace the strategy  $\tau_r$  for  $r$  in  $\tau$  by  $\sigma_r$ .  $\square$

A *social choice function* takes a preference profile and returns the desired outcomes. An *implementation* of such a function is a mechanism that is set up in such a way that the best strategies – according to some solution concept taken from game theory proper, such as Nash Equilibrium, or survival after iterated elimination – for the individual players gives the desired

result. The idea is that the mechanism implements the function in the sense that it is in player’s best interest to behave in such a way that the intended result is obtained.

**Definition 3 (Implementation)** We say that a protocol  $P$  implements a certain property  $X$  for a role specification  $\Theta$  just in case for each strategy profile  $\theta \in \Theta$  and each Nash Equilibrium  $\sigma$ , the play generated by  $\sigma(\theta)$  has the property  $X$ .  $\square$

Now, usually, mechanism design is concerned with finding protocols that are ‘robust’ in the sense that they implement a given social choice function within large classes of preference profiles. In our dialogue system above, we take a somewhat orthogonal approach: given a relatively unconstrained mechanism, we try to relate assumptions about preferences of the players with desired properties over the outcomes.

### 3 Persuasion Dialogue

In this section, we apply our approach to persuasion dialogues. A *persuasion* is characterized by a conflict of opinions between the *proponent* and the *opponent* of a claim. The purpose of the participants is to resolve the conflict by persuading the other party to give up their opinions. Protocols, or *dialogue systems*, for argumentation, persuasion or debate have been discussed extensively, originally in the literature on rhetoric and law (Hamblin, 1970; Mackenzie, 1979; Walton and Krabbe, 1995), but increasingly also by researchers interested in agent communication (Parsons et al., 1998; McBurney and Parsons, 2002). For a recent review of formal dialogue systems for persuasion, we refer to Prakken (2006).

#### 3.1 Mackenzie’s dialogue system

We take the existing argumentation system of Mackenzie (1979), and show how some of its structural dialogue rules can be replaced by assumptions on the player’s preferences. We selected Mackenzie because it is a ‘classic’ in the field of argumentation theory. The system contains explicit structural rules, and can therefore be seen as a relatively restricted protocol. This makes it suitable for our experiment.

Mackenzie (1979) defines rules for a dialogue game in which two participants have a limited number of locutions available: statements, questions, answers, challenges, withdrawals and a kind of defense move. MacKenzie’s definitions are two-fold: he defines the effect that each locution has on the set of *commitments* of each of the players, and he defines a number

of *rules* that specify at which point a certain locution may or may not be used.

Unlike many contemporary dialogue systems for persuasion, statements in Mackenzie’s system do not only affect the commitments of the speaker, but also – implicitly – those of the hearer. This is called the principle of ‘silence means consent’. The hearer can only avoid making such tacit commitments by challenging the speaker, or explicitly withdrawing the commitment. Moreover, the system is designed to avoid the fallacy of “begging the question”, which occurs when (i) in order to know some premiss of the argument, one must already know the conclusion, or (ii) when one participant asks the other to grant him a premiss which involves the topic under dispute. According to an earlier paper by ?, begging the question could only be avoided by requiring dialogues to be ‘cumulative’: once a proposition has been established, it remains so. Mackenzie’s dialogue system is designed to show them wrong: it is non-cumulative, but avoids begging the question.

Although we try to stay as close as possible to MacKenzie’s original system, we have adapted his definitions. Our aim here is to show that, if we add certain assumptions about the preferences of the dialogue participants, we can make do without the rules, and *derive* them instead as properties of rational behavior. In other words, we show how many of MacKenzie’s rules can be *implemented* in a relatively liberal protocol together with a specification of a number of reasonable constraints on the preferences of the players.

#### 3.2 Dialogue Protocol

In the *dialogue protocol* that we are defining, two players, *Ann* and *Bob*, alternately choose from a limited number of *locutions*. *Ann* starts. Moreover, at each point in the dialogue a participant can, if it is his turn, *terminate* the dialogue. A locution is a speech act with a propositional content  $\varphi$  expressed in propositional logic. The speech act can be either a *statement*, a *question* or a *discommitment*.

**Definition 4 (Locutions)** For  $\varphi$  a sentence of propositional logic, the set of locutions  $L$  is given by:

$$l ::= \text{state}(\varphi) \mid \text{question}(\varphi) \mid \text{discommit}(\varphi)$$

We will call any sequence of locutions generated by this protocol a *dialogue*. It will be useful to represent a dialogue as a sequence of pairs from  $\{\text{Ann}, \text{Bob}\} \times L$ . We write  $(i, l)_k$  for the  $k$ -th turn in a dialogue.

With each turn  $k$  we associate in the dialogue the sets  $C_k^A$ ,  $D_k^A$ ,  $C_k^B$  and  $D_k^B$  – the *commitments* and *doubts*

locution	effect on speaker (A)	effect on hearer (B)
<i>statement</i> (A, state( $\varphi$ )) <sub>k</sub> and not (B, discommit( $\psi$ )) <sub>k-1</sub> and not (B, question( $\varphi$ )) <sub>k-1</sub>	$C_{k+1}^A = C_k^A \cup \{\varphi\}$ $D_{k+1}^A = D_k^A$	$C_{k+1}^B = C_k^B \cup \{\varphi\}$ $D_{k+1}^B = D_k^B \setminus \{\varphi\}$
<i>withdrawal, challenge:</i> (A, discommit( $\varphi$ )) <sub>k</sub>	$C_{k+1}^A = C_k^A \setminus \{\varphi\}$ $D_{k+1}^A = D_k^A \cup \{\varphi\}$	$C_{k+1}^B = C_k^B$ $D_{k+1}^B = D_k^B \cup \{\varphi\}$
<i>defense:</i> (A, state( $\varphi$ )) <sub>k</sub> and (B, discommit( $\psi$ )) <sub>k-1</sub>	$C_{k+1}^A = C_k^A \cup \{\varphi, \varphi \rightarrow \psi\}$ $D_{k+1}^A = D_k^A \setminus \{\psi\}$	$C_{k+1}^B = C_k^B \cup \{\varphi, \varphi \rightarrow \psi\}$ $D_{k+1}^B = D_k^B$
<i>question:</i> (A, question( $\varphi$ )) <sub>k</sub>	$C_{k+1}^A = C_k^A$ $D_{k+1}^A = D_k^A \cup \{\varphi, \neg\varphi\}$	$C_{k+1}^B = C_k^B$ $D_{k+1}^B = D_k^B \cup \{\varphi, \neg\varphi\}$
<i>answer:</i> (A, state( $\varphi$ )) <sub>k</sub> and (B, question( $\psi$ )) <sub>k-1</sub> (with $\varphi \in \{\psi, \neg\psi\}$ )	$C_{k+1}^A = C_k^A \cup \{\varphi\}$ $D_{k+1}^A = D_k^A \setminus \{\varphi\}$	$C_{k+1}^B = C_k^B \cup \{\varphi\}$ $D_{k+1}^B = D_k^B \setminus \varphi$

Table 1: Adapted Dialogue System of MacKenzie

of, respectively, Ann and Bill, based on the locutions they have made up to that point. Intuitively, the commitments of a player are those sentences that he publicly expressed to believe, while the doubts are those facts that he professed to question or doubt<sup>1</sup>. The dialogue rules that regulate the effect of locutions on commitments and doubt are presented in table 1.

Our protocol only has three types of basic locutions. The exact effects of these locutions on the commitments of speaker and hearer depend on the context in which they are made. The locutions from Mackenzie’s dialogue system can be expressed in terms of these primitives, given the dialogue context:

- Whenever a locution  $\text{state}(\varphi)$  does not immediately follow a question or a discommitment, it is called a *statement*. The result is that both speaker and hearer are committed to the truth of its content  $\varphi$ . If  $\varphi$  was doubted by the hearer earlier, it is now retracted from her doubts.

The fact that a statement not only affects the speaker’s commitments to  $\varphi$ , but also those of the hearer, is meant to model the principle of ‘silence means consent’. The hearer can avoid making a tacit commitment, by immediately following up with a challenge or withdrawal.

- In a context in which the speaker is committed to  $\varphi$ , a locution  $\text{discommit}(\varphi)$  is called a *withdrawal*. It moves  $\varphi$  from the commitments of the speaker

to his doubts, and moves  $\varphi$  to the doubts of the hearer as well.

- In a context in which  $\varphi$  is a part of the commitments of the hearer and not of the speaker,  $\text{discommit}(\varphi)$  is called a *challenge*.
- A locution  $\text{state}(\varphi)$  is called a *defense* whenever it immediately follows a  $\text{discommit}(\psi)$  by the other participant. The effect is that  $\varphi$  is interpreted as providing an argument or a motivation for  $\psi$ : the speaker will be committed to both  $\varphi$  and  $\varphi \rightarrow \psi$ . Consequently, the challenged proposition  $\psi$  is now considered to be defended and can be removed from the doubts.
- A *question*  $\text{question}(\varphi)$  puts both  $\varphi$  and  $\neg\varphi$  in the doubts of the speaker, indicating that an answer to the question is not known.
- A location  $\text{state}(\varphi)$  is called an *answer* if it immediately follows a question. An answer commits both speaker and hearer to the truth of  $\varphi$ , and removes  $\varphi$  from the doubts of either of them.
- A *demand for resolution* makes use of the relation of *immediate consequence* that is defined below. A locution by A of the form  $\text{state}(\Phi \rightarrow \psi)$  at time  $k$  is called a *demand for resolution* just in case (1)  $\Phi \xrightarrow{\text{imc}} \psi$  and (2) the commitments ( $C_B, D_B$ ) at time  $k$  are such that  $\Phi \subseteq C_B$  and  $\neg\psi \in C_B$  or  $\psi \in D_B$ .

<sup>1</sup>In Mackenzie, challenges are represented as special commitments of the form ( $i, \text{why}(\varphi)$ ). Here we use a separate set of doubts.

The main difference with MacKenzie’s system is that in our dialogue protocol, the players are at complete

liberty to choose the locutions they want, and can exit at any moment. This fits in with our policy of leaving the game structure as liberal as possible, and deriving the (in)felicity of dialogue moves from the preferences of the participants.

### 3.3 Role Specification

We can now proceed to define a role specification that embodies natural assumptions about participants in persuasion dialogues, to predict rational behavior in the resulting dialogue game. We have a wide range of reasonable choices here. In antagonistic settings, players' interests are opposed – for example, in a political debate, participants may try to convince the other of opposing views. This contrasts with more cooperative games, in which the interests of the participants are aligned, for example in a situation in which participants try to learn from each other. We may even consider absurdist Ionesco-type settings, in which players say more or less arbitrary, unconnected things – dialogues in which players do not have very much interest in the coherence of their utterances, and in which they are, perhaps, more interested in the aesthetics of the resulting sequences of locutions.

Absurdism aside, we will consider a relatively weak role specification for a more or less antagonistic setting: we assume that players do not want to contradict themselves, and that, all things being equal, they prefer to catch the other participants in a contradiction rather than not. Moreover, all else being equal, they prefer finite dialogues over infinite ones.

For the purpose of this particular example, we will assume with MacKenzie that participants have very limited logical capabilities – in particular, they do not oversee all the logical consequences of their commitments. To capture this, MacKenzie introduces a simple relation  $\Phi \xrightarrow{imc} \varphi$  between sets of sentences and single sentence, called *immediate consequence*. MacKenzie remains vague about the details, and so will we, but he reasonably assumes that it should include at least direct applications of modus ponens: if  $\varphi \in \Phi$  and  $\varphi \rightarrow \psi \in \Phi$ , then  $\Phi \xrightarrow{imc} \psi$ . We will assume it includes identity as well. The idea is that ‘immediate consequence’ represents the minimal logic that agents are aware of: immediate consequence conditionals are undeniably true even for imperfect reasoners.

**Definition 5 (immediate consequence)** A pair  $(C, D)$  is said to be immediately inconsistent iff there is a finite  $\Phi \subseteq C$  such that  $\Phi \rightarrow \psi \in C$ ,  $\Phi \xrightarrow{imc} \psi$ , and  $\neg\psi \in C$  or  $\psi \in D$ .

Let  $d$  be a dialogue in our dialogue game of length  $k$ ,

and suppose the last move was the termination of the dialogue by player  $A$ . We say that the *outcome commitments* of the other player,  $B$ , are his commitments just before the termination, i.e. at move  $k - 1$ , while the outcome commitments of  $A$  are those from before that move, at state  $k - 2$ . When  $k = 1$ , we just let these commitments be empty. Less formally, in a finite dialogue terminated by  $A$ , the commitments of  $B$  are determined by all that has been said, while  $A$ 's commitments are determined by all locutions except the last locution by  $B$ .

We can now define a possible role specification: players try to avoid outcome commitments which are immediately inconsistent, but, all things being equal, prefer those of the other player to be just that.

1. A player strictly prefers a dialogue in which her outcome commitments are not immediately inconsistent but her opponents are over other outcomes (we say that in this case she *wins* the dialogue).
2. Among the dialogues that she wins, a player prefers shorter dialogues over longer ones.
3. A player *loses* a dialogue if her outcome commitments are immediately inconsistent. She prefers any other outcome over losing the dialogue.

Of course, these assumptions are arbitrary, up to a point. We have chosen them because they have some intuitive attraction, and are strong enough to allow us to derive a large part of MacKenzie's dialogue rules as implemented properties.

**Proposition 1 (Derived Dialogue Rules)** The protocol and the role specifications of the previous sections implement the following properties, in the sense that the following properties are true for all dialogues generated by a Nash Equilibrium:

- *Resolve personal inconsistencies immediately.* In each dialogue, if it is  $A$ 's turn and her commitments are immediately inconsistent, then she will (and can) choose a locution that removes this inconsistency.
- *An immediate consequence condition may not be withdrawn* (MacKenzie's rule  $R_{Imcon}$ )
- *An immediate consequence condition may not be challenged* (MacKenzie's rule  $R_{LogChall}$ )
- *Answer questions immediately,* or be committed to having the content of the question among your doubts. (weaker version of MacKenzie's  $R_{Quest}$ )

- *Respond to a challenge* by either withdrawing the challenged proposition, or by defending it with an argument that is consistent with your previous commitments. (stronger version of MacKenzie’s  $R_{\text{Chall}}$ )
- *Respond to a demand for resolution* by either withdrawing one of its premisses, or its conclusion ( $R_{\text{resolve}}$  in MacKenzie.)

*Proof:* Precise proofs of these claims (and precise formulations of them) can be found in the appendix.  $\square$

This shows that many of the crucial properties of the dialogues generated by Mackenzie’s structural dialogue rules, can now be derived. Because of the ‘silence means consent’ principle, a statement that goes unchallenged, implicitly also commits the hearer to its truth. To avoid this, the hearer may challenge the statement, or explicitly withdraw the commitment. Typically, this will lead to a long chain of challenges, followed by defense moves, and further challenges, until either of the players gives up. For Mackenzie, it is important that the resulting dialogues avoid ‘begging the question’. To prevent this Mackenzie mainly uses  $R_{\text{Chall}}$ , which originally states that after a challenge by  $A$ ,  $B$  must respond by (i) a withdrawal, (ii) a demand for resolution, or (iii) a defense move. This characteristic is preserved in our version. Our version of the system also remains non-cumulative in this sense: it is always possible to withdraw a previously made commitment.

## 4 Related Research

There are three kinds of research related to our cross-over of game theory and dialogue design.

### 4.1 About Flexible Interaction

Our work relates to research in the multi-agents systems community, which is looking for more open or flexible approaches to agent interaction (Mazouzi et al., 2002; Yolum and Singh, 2002). There are two ways in which an interaction protocol can be made more flexible: by the designer, off-line, or by the participants themselves, on-line.

We expect that techniques from the Semantic Web community, like automated search facilities and tools for combining web-services (Antoniou and van Harmelen, 2004), will make it increasingly possible for automated agents to establish coordination rules themselves. On the Semantic Web, the interaction protocols and the activities that they regulate are typically offered in conjunction. Thus, detailed information is

available about the context of application, which warrants the kinds of assumptions about preferences of participants that we want to make.

### 4.2 About Argumentation and Persuasion

Argumentation provides an interesting case study for our approach. First, because in argumentation theory – like for us – the focus is not so much on the outcome of a debate, but rather on the arguments that are put forward during the debate, and their structure. Second, because argumentation techniques may also be applied to other kinds of interaction, like negotiation (Parsons et al., 1998). Thirdly, because we observe a similar tendency to make assumptions about dialogue participants. For example, Parsons et al. (2003) distinguish the following agent types: a *confident* agent can assert any proposition for which he can construct an argument, a *careful* agent can do so only if cannot construct a stronger counterargument (in terms of some measure of strength between arguments), and a *thoughtful* agent can do so only if he can construct an acceptable argument for the proposition (in terms of the underlying inference relation). Based on the agent type, Parsons et al. prove whether or not a move is rational, in a certain dialogue. We differ from this approach. To prove rationality, Parsons et al. need access to the individual belief bases of the agents. We, on the other hand, only make general assumptions, based on conventions of the dialogue type and the role of the participant.

### 4.3 About Game Theory and Dialogue

Even if we believe that our work uncovers a new way of combining mechanism design with dialogue games, we are, of course not the only ones to use game theory techniques for the study of dialogue. Here, we very shortly discuss some related work in this field.

Parikh (1991), who uses game theory to study the Gricean maxims for cooperative dialogue (Grice, 1975). In particular, Parikh develops an approach that combines game theory and situation theory to obtain a framework in which it is possible to model the strategic considerations that lead a speaker to choose a certain utterance and the hearer to interpret that utterance in a certain way. This provides a way to explain how the a speaker can, in a particular context, convey more information beyond that contained in the literal meaning of an utterance. Also (Gerbrandy, 2007) is related here. The main difference with our approach in this paper is that we are not so much concerned with the pragmatics of single utterances, but rather with the structure of a rather abstract type of dialogue game.

Another branch of research on persuasion using game

theoretic tools, is carried out by Glazer and Rubinstein (2004, 2006). They study a restricted scenario of information asymmetry, in which a speaker tries to persuade a listener to take a certain action, which is claimed to be beneficial to both. Now the listener's preferred action depends on the agent type of the speaker, which is known only to the speaker himself. The speaker therefore sends a message to the listener about his type. The listener can only find out the truth about one of the aspects that determine the type of the speaker. The objective for the designer of the game, is to determine which rules the listener should follow in order to avoid selecting the wrong action as much as possible.

There are several mechanisms. The listener can ask about one specific aspect deterministically, or select one at random to check. Glazer and Rubinstein (2004) show that selecting the optimal mechanism comes down to solving a linear programming problem to minimize mistakes. An optimal mechanism always exists, partly randomized, and only under certain assumptions, a deterministic mechanism can be shown to exist. Glazer and Rubinstein (2006) show for a slightly different scenario, that any optimal persuasion rule is also ex-post optimal, which is quite remarkable. However, as a general analysis of persuasion dialogue, the scenario is somewhat restricted. But as Glazer and Rubinstein admit, their purpose is “not to suggest a general theory for the pragmatics of persuasion, but rather to demonstrate a rationale for inferences in persuasion situations” (Glazer and Rubinstein, 2006).

## 5 Conclusions

Participants in an interaction can be seen as players in a game, specified by a protocol. Based on conventions of the interaction type and the roles of the players, we can make assumptions about their expectations and references. Using such assumptions, we can use game theoretic techniques to predict (up to a point) what a rational player will do. This observation gives us some leeway when defining a protocol. Instead of having to specify in detail what moves players are allowed to make at each point, we can replace the protocol by a more liberal one that specifies only partially what role players must do. One could say that we move the responsibility for the desired outcome of the protocol away from the designer and towards the participants.

Whether this is useful or not depends on the application, of course. A protocol for preventing nuclear meltdown should perhaps not be made more liberal, relying on the rationality of the players. Such a protocol needs to be foolproof: safe even when the players are not completely rational. Similarly, protocols that

are designed to implement a technical convention, like the TCP-IP protocol, cannot be generated by rationality assumptions. However, if a protocol is meant to streamline communication flows in a particular context, a designer may be better off by not defining that flow explicitly. Instead, a limited set of locutions and syntactic and semantic agreements may be sufficient. Smart role players may even improve on the protocol rules by just choosing the best action, so that explicit rules might turn out to be a hindrance.

The point of this paper is to illustrate, by example, how such a flexible approach to interaction design might work, and which kind of formal tools can be used for proving properties of the resulting ‘liberal’ protocols.

## Appendix: Proofs

We show that the protocol we defined in this paper indeed implements the properties of Proposition 1.

The Role Specification of the participants – that is, the minimal constraints on their preferences – must be the following:

- A player *wins* a dialogue iff her final commitments are not immediately inconsistent and the other player’s final commitments are immediately inconsistent. A player absolutely prefers those dialogues that she wins.
- Among those dialogues that she wins, a player prefers shorter dialogues over longer ones.
- A player *loses* a dialogue iff her final commitments are immediately inconsistent. A player least prefers those outcomes in which she loses.

**Lemma 1 (always immediately resolve personal inconsistencies)** In each dialogue, if  $(C, D)_k^A$  is immediately inconsistent, and it is  $A$ ’s turn, then  $(C, D)_{k+1}^A$  is consistent (and the same holds for  $B$ ).

*proof:* We prove the desired result by induction on  $k$ . Suppose  $(\sigma_A, \sigma_B)$  is in Nash Equilibrium, and generates a dialogue of length  $k$  in which  $(C, D)_k^A$  is immediately inconsistent. We need to show that if this is the case, it is  $A$ ’s turn, and at the next turn, her commitments are consistent.

If  $k = 0$ , this holds trivially, as the antecedent is false.

So, assume  $k > 0$ . By induction hypothesis, we know that  $(C, D)_{k-1}^A$  is consistent (because it was  $A$ ’s turn at  $k-2$ ). Also  $(C, D)_k^B$  must be consistent by induction hypothesis.

Now, suppose that  $A$  chooses a locution that renders her own state inconsistent.  $B$  has the option to exit at  $k + 1$ . Since this would lead to a final outcome in which  $B$ 's commitments are consistent, but  $A$ 's is not, it is actually in  $B$ 's advantage to do just that. Given our assumptions,  $B$  prefers winning immediately over winning later, so if  $\sigma$  is a Nash Equilibrium, then  $\sigma_B$  will prescribe to exit at this point  $k + 1$ .

But this outcome – in which  $A$ 's commitments are inconsistent – is absolutely dispreferred by  $A$ ; she is better off choosing a different locution at  $k$  (or to exit). Which means that  $\sigma_A$  is not a Nash Equilibrium.  $\square$

We now are ready to prove some of MacKenzie's rules:

**Lemma 2**  $R_{\text{Imcon}}$ : A conditional whose consequent is an immediate consequence of its antecedent must not be withdrawn.

*proof:* Using lemma 1, because this withdrawal would render  $A$ 's state immediately inconsistent, and this move is excluded in any rational dialogue.  $\square$

**Lemma 3**  $R_{\text{LogChall}}$ : A conditional whose consequent is an immediate consequence of its antecedent must not be challenged.

*proof:* This immediately follows from lemma 1.

**Lemma 4 (Answer questions)** Answer a question  $\text{question}(\varphi)$  immediately by either stating  $\varphi$  or its negation, or by withdrawing either  $\varphi$  or its negation (when appropriate). Otherwise, be committed to having the content of the question among your doubts. (This is a weaker form of MacKenzie's  $R_{\text{Quest}}$ , which reads: After "Is the case that  $P$ ?", the next event must either be ' $P$ ', ' $\text{Not } P$ ', or ' $\text{Withdraw } P$ '.)

*proof:* This observation is not so much a game-theoretic consequence of our solution concept, but rather an observation about the way we have defined the commitment rules. Suppose  $A$  has asked the question whether  $\varphi$ . Now, we can distinguish three cases: (1)  $B$  was already committed to  $\varphi$  (or to  $\neg\varphi$ ). In this case his commitments are immediately inconsistent, and he *must* (by lemma 1) resolve this inconsistency by either stating that  $\varphi$  or  $\neg\varphi$  and thereby removing  $\varphi$  or  $\neg\varphi$  from his doubts, or by withdrawing his commitment to  $\varphi$  (or to  $\neg\varphi$ ).

(2)  $B$  is not committed to  $\varphi$  or  $\neg\varphi$ , and answers the question; and

(3)  $B$  continues with an unrelated locution. In this case,  $B$  will have both  $\varphi$  and  $\neg\varphi$  among his Doubts.  $\square$

**Lemma 5 (Respond to a challenge with a withdrawal, or a defense)**

A challenge by  $A$  with  $\varphi$  is followed by either:

- (1) a withdrawal by  $B$  of the challenged  $\varphi$ , or
- (2) giving a defending argument for it:  $(B, \text{state}(\psi))$  for some  $\psi \notin D_k^A$ .

(This is a strong variant of MacKenzie's  $R_{\text{Chall}}$ , which reads: "After Why  $P$ ?, the next event must be either (i) No commitment  $P$ ; or (ii) The resolution demand of an immediate consequence conditional whose consequent is  $P$  and whose antecedent is a conjunction of statements to which the challenger is committed; or (iii) A statement not under challenge with respect to its speaker, i.e., a statement to whose challenge its hearer is not committed.)

*proof:* Remember that  $(A : \text{discommit}(\varphi))_k$  is a challenge just in case  $\varphi \in C_k^B$ .

This results in a state such that  $\varphi \in C_{k+1}^B$  and  $\varphi \in D_{k+1}^B$ , which is an immediate inconsistency in  $B$ 's commitments.

We know from lemma 1 that  $B$  will resolve this with his next utterance.  $B$  has the choice between two locutions to resolve this conflict. He can give a *defense* by using  $\text{state}(\psi)$  for some  $\psi$  that is consistent with his commitments (and in particular, with his doubts). This resolves the conflict, by withdrawing  $\varphi$  from his commitments. (The new assertion must be consistent with his commitments up till now, again by 1.) Alternatively,  $B$  may *withdraw*  $\psi$ .  $\square$

**Lemma 6 (withdraw premiss or conclusion after a demand for resolution)**

After a demand for resolution of  $\varphi \rightarrow \psi$  by  $A$ ,  $B$  must either withdraw or challenge one of the premisses  $\varphi$  or the sentence  $\neg\psi$ .

(This is our version of Mackenzies rule  $R_{\text{Resolve}}$ )

*proof:* Recall the definition of a demand for resolution. It holds  $(A : \Phi \rightarrow \psi)_k$ , and that  $\Phi \xrightarrow{\text{imc}} \psi$ , that  $\Phi \subseteq C_k^B$  and  $\psi \in D_k^B$  or  $\neg\psi \in C_k^B$ .

Now,  $C_{k+1}^B$  is immediately inconsistent, and we know, by lemma 1 that  $B$  must resolve this inconsistency at  $k + 2$ . He cannot do this by discommitting  $\Phi \rightarrow \psi$ , because it is an immediate consequence, and it would leave his commitments inconsistent. So he must resolve the inconsistency by discommitting (that is, challenge or withdraw) one of the premisses of  $\Phi$ , or, if the inconsistency arises because  $\neg\psi \in C$  (and  $\psi \notin I$ ), by discommitting  $\neg\psi$ .  $\square$

## References

- Antoniou, G. and F. van Harmelen: 2004, *A Semantic Web Primer*. MIT Press.
- Gerbrandy, J.: 2007, 'Communication Strategies in Games'. *Journal of Applied Non-Classical Logics* **17**(2).
- Glazer, J. and A. Rubinstein: 2004, 'On Optimal Rules of Persuasion'. *Econometrica* **72**, 1715–1736.
- Glazer, J. and A. Rubinstein: 2006, 'On the Pragmatics of Persuasion: A Game Theoretical Approach'. *Theoretical Economics* **1**, 395–410.
- Grice, H. P.: 1975, 'Logic and Conversation'. In: P. Cole and J. L. Morgan (eds.): *Syntax and Semantics 3*. Academic Press, New York, pp. 41–58.
- Hamblin, C. L.: 1970, *Fallacies*. Methuen & Co, London.
- Mackenzie, J.: 1979, 'Question Begging in Non-cumulative Systems'. *Journal of Philosophical Logic* **8**, 117–133.
- Mazouzi, H., A. E. F. Seghrouchni, and S. Haddad: 2002, 'Open protocol design for complex interactions in multi-agent systems'. In: *Proceedings of the 1st international joint conference on Autonomous agents and multiagent systems (AAMAS'02)*. pp. 517 – 526, ACM Press.
- McBurney, P. and S. Parsons: 2002, 'Games that agents play: A formal framework for dialogues between autonomous agents'. *Journal of Logic, Language and Information* **11**(3), 315–334.
- Parikh, P.: 1991, 'Communication and Strategic Inference'. *Linguistics and Philosophy* **14**, 473–514.
- Parsons, S., C. Sierra, and N. Jennings: 1998, 'Agents that reason and negotiate by arguing'. *Journal of Logic and Computation* **8**, 261–292.
- Parsons, S., M. Wooldridge, and L. Amgoud: 2003, 'Properties and complexity of some formal inter-agent dialogues'. *Journal of Logic and Computation* **13**, 347–376.
- Prakken, H.: 2006, 'Formal systems for persuasion dialogue'. *The Knowledge Engineering Review* **21**, 163–188.
- Walton, D. N. and E. C. Krabbe: 1995, *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press.
- Yolum, P. and M. P. Singh: 2002, 'Flexible Protocol Specification and Execution: Applying Event Calculus Planning using Commitments'. In: *Proceedings of the 1st International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS'02)*. pp. 527–534, ACM Press.