

# Improper behaviour in argumentation based persuasion dialogues

---

FACULTY OF HUMANITIES  
DEPARTMENT OF PHILOSOPHY  
COGNITIVE ARTIFICIAL INTELLIGENCE

THESIS FOR THE DEGREE OF MASTER OF SCIENCE

45 ECTS

*Author:*  
Rutger HOMMES

*Supervisor:*  
Prof. dr. Henry PRAKKEN

June 28th 2015

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Dialogue systems, persuasion and improper behaviour</b>	<b>6</b>
2.1	Introduction . . . . .	6
2.2	Concepts . . . . .	6
<b>3</b>	<b>Prakken’s framework</b>	<b>8</b>
3.1	Introduction . . . . .	8
3.2	Dung argumentation frameworks . . . . .	9
3.3	<i>ASPIC</i> <sup>+</sup> . . . . .	11
3.4	Prakken’s framework formally defined . . . . .	19
3.5	Liberal dialogue systems . . . . .	22
<b>4</b>	<b>Desirable properties of dialogue agents</b>	<b>24</b>
4.1	Dialogue participants . . . . .	25
4.2	Grice’s maxims . . . . .	25
<b>5</b>	<b>Improper behaviour of agents in liberal dialogues</b>	<b>28</b>
5.1	Preliminary definitions . . . . .	29
5.2	Dishonesty . . . . .	35
5.2.1	Lying . . . . .	35
5.2.2	Bullshitting . . . . .	37
5.2.3	Dialogical incoherence . . . . .	38
5.2.4	Obvious logical incoherence . . . . .	42
5.2.5	Unobvious persuasiveness revisited . . . . .	43
5.3	Irrelevance . . . . .	45
5.4	Verbosity . . . . .	47
5.5	Improper behaviour unrelated to individual moves . . . . .	47
<b>6</b>	<b>Variation 1: inquiry</b>	<b>48</b>
6.1	Extension of the communication language . . . . .	49
6.2	Improper behaviour related to inquiry . . . . .	49
6.2.1	Pretending to be ignorant . . . . .	49
6.2.2	Asking an irrelevant question . . . . .	51
6.2.3	Asking the same question more than once . . . . .	53
<b>7</b>	<b>Variation 2: indications of incoherence</b>	<b>53</b>
7.1	Exploiting logical incoherence: <i>resolve</i> $\varphi, -\varphi$ . . . . .	53
7.2	Exploiting dialogical incoherence: <i>eo ipso</i> $\varphi$ . . . . .	54

<b>8</b>	<b>Definitions of improper behaviour</b>	<b>56</b>
8.1	Definitions without non-optimal usefulness . . . . .	56
8.2	Defining different types of non-optimal usefulness . . . . .	57
8.2.1	Premature surrendering . . . . .	57
8.2.2	Attacking with a non-most persuasive argument . . . . .	58
8.3	Definition including non-optimal usefulness . . . . .	59
<b>9</b>	<b>Banning improper behaviour from persuasion dialogues</b>	<b>59</b>
9.1	Guidelines for agent design . . . . .	60
9.2	Protocol rules . . . . .	61
9.3	Justification of the instruments for ensuring proper behaviour	62
<b>10</b>	<b>Conclusion</b>	<b>64</b>

## **Abstract**

In this thesis, it is investigated how agents participating in argumentation based persuasion dialogues should behave ideally. Also, types of improper behaviour of agents participating in such dialogues are defined, categorized and discussed. In addition, protocol rules and guidelines for agent design are devised using which improper agent behaviour can be banned from persuasion dialogues.

# 1 Introduction

In this thesis, improper behaviour of agents participating in argumentation based persuasion dialogues is studied. Such dialogues are modelled by dialogue systems, which define the principles of a coherent dialogue in which the participants have a particular goal. In the case of this thesis, one participant's goal is to persuade a second participant of some claim or argument. Hence we call the dialogues discussed in this thesis '*persuasion dialogues*', a term which was coined in [12]. We also call these dialogues '*argumentation based*' since the persuading participant of a dialogue attempts to persuade the other participant by means of arguments, to which that other participant can respond with counterarguments, which, again, can be attacked by the first participant using other counterarguments, and so on.

The following is a typical argumentation based persuasion dialogue.

**Arie:** Taking a cold shower is healthy. (making a claim)

**Bert:** Why is taking a cold shower healthy? (asking grounds for a claim)

**Arie:** Since it stimulates one's blood circulation. (giving an argument for a claim)

**Bert:** It is true that taking a cold shower stimulates one's blood circulation (conceding a claim), but I disagree with you on the fact that this makes taking a cold shower healthy, since I read on Wikipedia that the stimulation takes place in such a way that it can lead to fainting, which is not healthy. (providing a counterargument)

**Arie:** Indeed, that is what Wikipedia says (conceding a claim) and I agree that fainting is not healthy (conceding another claim), but this does not prove anything, since Wikipedia is an unreliable source of information. (undercutting a counterargument)

**Bert:** That is a good point, one should not blindly trust Wikipedia (conceding a claim), but still taking a cold shower is not healthy, since it is usually a very unpleasant experience. (offering an alternative counterargument)

Yet, the fact that a dialogue system models dialogues with some degree of coherence, such as the example dialogue above, does not guarantee that the participants of these dialogues behave properly. Indeed, participants can display *improper behaviour*. An important class of such behaviours is discussed in [1], where it is called 'dishonesty' and a distinction is made between lying, bullshitting (telling something without having the proper evidence for it) and deception. Whereas in that research general notions of these types of dishonesty are introduced, definitions for these types of dishonesty which are designed specifically for persuasion will be presented in this thesis.

Furthermore, given a dialogue system which models dialogues from which particular aspects of incoherence are banned, participants can still display behaviour which results in other aspects of incoherence. In [8], different dialogue systems, which are instantiations of a general framework that we will call “Prakken’s framework” from here on, are defined in which the actions of agents, which are more or less relevant to the dialogue depending on what definition of relevance is adopted, can cause different types of incoherence. In this thesis, the different notions of relevance discussed in [8] are linked to the concept of improper behaviour.

The following questions will be answered in this thesis:

- How should participants of persuasion dialogues behave ideally?
- Depending on the design choices made, what types of improper behaviour of dialogue participants can occur in Prakken’s framework?
- How can these types of improper behaviour be banned from dialogues which are modelled by Prakken’s framework?

Important fields of research within AI to which the results of this thesis contribute, are automated reasoning and agent technology. Studying argumentation based dialogues is important to the field of automated reasoning, since human beings themselves use argumentation both in their individual reasoning and in communication with others. It is, therefore, likely that a thorough understanding of argumentation is necessary in order for science to make advances towards the achievement of “hard AI”, i.e. the creation of machines with reasoning capabilities similar to those of human beings. The research presented in this thesis contributes to the field of automated reasoning, since it builds upon research conducted in the area of argumentation based persuasion dialogues. Also, agent technology is an important field in AI, because agents are autonomous entities which possess AI to some degree. In order for these agents to be able to function in a human-like way, they need to be given the ability to communicate. This communication can have different goals, one of which is persuasion. The research presented in this thesis contributes to the field of agent technology, since it treats realistic argumentation based persuasion dialogues in which agents can display improper behaviour such as dishonesty.

In Section 2, argumentation based dialogue systems for persuasion are introduced. After this, we introduce and formally present Prakken’s framework in Section 3. In this section, we also make a number of design choices regarding instantiations of the framework. Next, we formally define the dialogue agents that will be participating in dialogues of these instantiations in

Section 4 and discuss ideal behaviour of dialogue agents using Grice’s conversational maxims. From this ideal behaviour, we move on to non-ideal, or improper, behaviour of agents in Section 5. We formally define, categorize and discuss different types of such behaviour in this section. As an extension of this section, we discuss, in Sections 6 and 7, even more types of improper behaviour which can occur in dialogue systems that are instantiations of Prakken’s framework under different design choices than the ones made in Section 3. In Section 9, we devise protocol rules and guidelines for agent design which can be used to ban the improper behaviours we discussed from a particular class of instantiations of Prakken’s framework. Finally, we answer our research questions in Section 10 and also give suggestions for further research in this section.

## **2 Dialogue systems, persuasion and improper behaviour**

### **2.1 Introduction**

In his influential paper, Douglas Walton [11] classified dialogues into six types according to their goal. One of these types is the so-called ‘persuasion dialogue’. In [12], it is said that the goal of a persuasion dialogue is to resolve a conflict of points of view between at least two participants by verbal means. In most of the literature, two-party persuasion dialogues have been investigated. In this thesis, we will also only look at persuasion dialogues with two participants: a proponent and an opponent of some proposition.

Besides the number of participants, there are two other design choices that are made in most of the literature on the subject, which we will also adopt:

- The commitments of the participants at the start of a dialogue should not conflict with their points of view.
- At most one side in a persuasion dialogue gives up.

### **2.2 Concepts**

In [9], Prakken gives an overview of the most important concepts on which dialogue systems are built. The following list of concepts is based on this overview:

- Dialogue systems have a *dialogue goal* (in the case of persuasion: reaching an agreement) and *at least two participants*, who can have various *roles* (in the case of persuasion: proponent and opponent).
- Each participant has a, possibly empty, set of *commitments*, which usually changes during a dialogue. For example: when an agent claims some proposition  $\varphi$ , then usually this means that the agent commits herself to  $\varphi$ .
- Dialogue systems have two languages, a *topic language* and a *communication language*. The communication language contains so-called *speech acts*, which are conversational actions. An example of a speech act in natural language is “promise”: when someone says “I promise  $x$ ” to someone else, then the act of uttering that sentence is the actual promising itself. An example of a speech act in persuasion dialogues is “claim”, where the claiming of  $x$  is actually done by the uttering of “claim  $x$ ” (we will encounter more speech acts later on).
- The distinction between *pure persuasion* and *conflict resolution* dialogues is made: the outcome of pure persuasion dialogues is fully determined by the participants’ points of view and commitments, while the outcome of conflict resolution is not.
- A dialogue system has *effect rules*, which specify the effects of utterances on the participants’ commitments, and *outcome rules*, defining the outcome of a dialogue.
- The participants of dialogues are *agents*. Each agent has her own *belief base*.
- A dialogue consists of *moves*. An example of a move is for an agent to give an argument for some proposition. Another example is for an agent to attack another agent’s argument by means of a counterargument.
- A dialogue *protocol* specifies the legal moves at each stage of a dialogue.
- A protocol consists of *rules*, which can be for a participant’s *dialogical* or *internal consistency*, for *dialogical coherence* or for *dialogical structure*.
- A protocol has a *public semantics* iff the legality of any move in a dialogue of a dialogue system using this protocol, can be verified by an outside observer.

- A protocol is *unique-move* if the turn shifts after each move; it is *multiple-move* otherwise.
- The *context* of a dialogue contains the knowledge that is presupposed and must be respected during the dialogue. The context is assumed consistent and remains the same throughout a dialogue.
- A protocol is *context-independent* if the set of legal moves and the outcome are always independent of the context.
- Furthermore, two kinds of protocol rules are sometimes separately defined: *turntaking* and *termination* rules.

In this thesis, propositional logic with a classical interpretation will be used as the topic language in our examples, unless explicitly stated otherwise.

## 3 Prakken’s framework

### 3.1 Introduction

In [8], a formal framework is presented for a very general class of two-party argumentation dialogues. The framework leaves a great deal of design choices open, since it allows for:

- different underlying logics,
- different sets of speech acts,
- varying degrees of coherence and flexibility in dialogues,
- different turntaking rules, and
- different rules on whether postponing of replies, multiple replies and coming back to earlier choices (backtracking) is allowed.

Yet, there is also a basic structure imposed on all dialogues in the study, which is an explicit reply structure on moves, where each move either *attacks* or *surrenders to* one earlier (but not necessarily the last) move of the other player. Another assumption of the framework is that during a dialogue the players implicitly build a structure of arguments and counterarguments related to the dialogue topic.

Although the framework leaves a great deal of freedom when it comes to the logical structure of the parties’ individual reasoning, some choices need

necessarily be made. One is the choice of a nonmonotonic logic. This is a type of logic in which, contrary to monotonic logics, when some conclusion  $c$  follows from a set of premises  $S$ , it is possible that when another premise is added to  $S$ ,  $c$  does not follow from  $S$  any more. In his framework, Prakken uses for this the *ASPIC*<sup>+</sup> formalism developed in [10]. The *ASPIC*<sup>+</sup> formalism builds, again, on Dung’s theory of abstract argumentation developed in [3].

### 3.2 Dung argumentation frameworks

The following definition of a Dung abstract argumentation framework is presented in [10].

**Definition 3.2.1** [Abstract argumentation framework] An *abstract argumentation framework* (AF) is a pair  $\langle \mathcal{A}, Def \rangle$ .  $\mathcal{A}$  is a set of arguments and  $Def \subseteq \mathcal{A} \times \mathcal{A}$  is a binary relation of defeat. We say that an argument  $A$  defeats an argument  $B$  iff  $(A, B) \in Def$ .

Arguments of an AF are assigned a status according to some given semantics: ‘justified’ if the argument is winning, ‘overruled’ if it is losing and ‘defensible’ if it is in a tie. Justified arguments are evaluated based on certain subsets of  $\mathcal{A}$  which are called *extensions*, a notion of which the definition depends on the chosen semantics (of which Dung gives a number of variants in his article). An extension consists of arguments which do not attack each other (extensions are *conflict-free*) and attack any argument that in turn attacks an argument in the extension (extensions *defend* or *reinstate* the arguments they contain). The following definitions for these concepts are presented in [10]:

**Definition 3.2.2** [Conflict-free, Defence] Let  $\mathcal{B} \subseteq \mathcal{A}$ .

- A set  $\mathcal{B}$  is *conflict-free* iff there exist no  $A_i, A_j$  in  $\mathcal{B}$  such that  $A_i$  defeats  $A_j$ .
- A set  $\mathcal{B}$  *defends* an argument  $A_i$  iff for each argument  $A_j \in \mathcal{A}$ , if  $A_j$  defeats  $A_i$ , then there exists  $A_k$  in  $\mathcal{B}$  such that  $A_k$  defeats  $A_j$ .

In [3], Dung presents definitions for different semantics for his argumentation frameworks. He defines not one but four different semantics, because for some argumentation frameworks, there are multiple intuitive ways to perform status assignment. However, there are also argumentation frameworks for which only one status assignment is intuitive, and indeed it is, as we will see,

the case that all of Dung’s different semantics yield the same unique status assignment for these frameworks.

Dung’s definitions are put together in the following single definition in [10]:

**Definition 3.2.3** [Acceptability semantics] Let  $\mathcal{B}$  be a conflict-free set of arguments, and let  $\mathcal{F} : 2^{\mathcal{A}} \mapsto 2^{\mathcal{A}}$  be a function such that  $\mathcal{F}(\mathcal{B}) = \{A \mid \mathcal{B} \text{ defends } A\}$ .

- $\mathcal{B}$  is *admissible* iff  $\mathcal{B} \subseteq \mathcal{F}(\mathcal{B})$ .
- $\mathcal{B}$  is a *complete extension* iff  $\mathcal{B} = \mathcal{F}(\mathcal{B})$ .
- $\mathcal{B}$  is a *grounded extension* iff it is the smallest (w.r.t. set-inclusion) complete extension.
- $\mathcal{B}$  is a *preferred extension* iff it is a maximal (w.r.t. set-inclusion) complete extension (or, equivalently, if  $\mathcal{B}$  is a maximal (w.r.t. set inclusion) admissible set).
- $\mathcal{B}$  is a *stable extension* iff it is a preferred extension that defeats all arguments in  $\mathcal{A} \setminus \mathcal{B}$ .

Suppose we have an argumentation framework with arguments  $A$ ,  $B$  and  $C$  and attack relation ‘ $A$  attacks  $B$  and  $B$  attacks  $C$ ’, which can be displayed in the form of a graph as follows:

$$A \longrightarrow B \longrightarrow C$$

When applied to this argumentation framework, which has an acyclic attack relation and a finite set of arguments, complete, grounded, preferred and stable semantics all yield  $\{A, C\}$  as the only extension. This extension, in fact, also corresponds to the only intuitive status assignment.  $A$  is clearly justified and therefore belongs in the extension, since it is not attacked by any argument.  $B$  is clearly overruled and does not belong in the extension, since its only attacker is  $A$ , which is, as we already established, a justified argument.  $C$ , again, is justified and belongs in the extension, not because it is not attacked at all, like  $A$ , but because it is attacked by only one argument which itself is overruled.

The distinctive behaviour of the different semantics becomes clear when they are applied to argumentation frameworks containing cyclic attack relations, such as the simple argumentation framework with arguments  $A$  and  $B$  and the cyclic attack relation:  $A$  attacks  $B$  and  $B$  attacks  $A$ , which can be

displayed by the following graph:

$$A \longleftrightarrow B$$

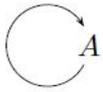
Given this argumentation framework, the extensions for the different kinds of the semantics defined by Dung are:

**Complete semantics**  $\emptyset$ ,  $\{A\}$  and  $\{B\}$ .

**Grounded semantics**  $\emptyset$ .

**Stable and preferred semantics**  $\{A\}$  and  $\{B\}$ .

The difference between stable and preferred extensions becomes clear when we apply them to an argumentation framework with only one argument,  $A$ , and the cyclic attack relation:  $A$  attacks  $A$ .



Now, the different extensions are:

**Stable semantics** none.

**Complete, grounded and preferred semantics**  $\emptyset$ .

### 3.3 $ASPIC^+$

The  $ASPIC^+$  framework combines two forms of reasoning within the abstract setting of argumentation defined by Dung in [3]: sound (i.e., deductive) reasoning on an uncertain basis and unsound (but still rational) reasoning on a solid basis.

With the exception of Definition 3.3.5, the following definitions are taken from [10].

**Definition 3.3.1** An  $ASPIC^+$  argumentation system is a triple  $AS = (L_t, R, n)$ , where:

- $L_t$ , (the *topic language*), is a logical language closed under negation ( $\neg$ ).

- $R = R_s \cup R_d$  is a set of strict ( $R_s$ ) and defeasible ( $R_d$ ) *inference rules* of the form  $\varphi_1, \dots, \varphi_n \rightarrow \varphi$  and  $\varphi_1, \dots, \varphi_n \Rightarrow \varphi$  respectively (where  $\varphi_i, \varphi$  are meta-variables ranging over well-formed formulas in  $L_t$ ), and  $R_s \cap R_d = \emptyset$ .
- $n$  is a *partial function* such that  $n : R_d \rightarrow L_t$ .

We write  $\psi = -\varphi$  just in case  $\psi = \neg\varphi$  or  $\varphi = \neg\psi$  (we will sometimes informally say that formulas  $\varphi$  and  $-\varphi$  are each other's negation).

The function  $n$  makes it possible to refer to defeasible inference rules in the topic language. Its usage becomes clear when we consider the following dialogue.

**Arie:** Donald Duck is not capable of rational thinking since he is a duck (stating a premise) and a duck is usually not capable of rational thinking. (using a defeasible inference rule)

**Bert:** It is true that Donald Duck is a duck (conceding a claim), but he is an exceptional duck that has the ability to speak and reason in the same way as human beings do, which means that he is, in fact, capable of rational thinking. (undercutting an argument)

The argument given by Bert in the example has as conclusion a proposition  $c$  with which he undercuts Arie's argument on a defeasible inference rule  $r$  that is used by Arie in his argument. In *ASPIC+*, this can be expressed by mapping  $r$  onto proposition  $n(r)$  in the topic language and making  $c = -n(r)$ , which negates the proposition.

The topic language of an argumentation system may contain propositions which actually do not have anything to do with the dialogue which one attempts to model. For example: in the example about Donald Duck, statements about the effects of a cold shower on one's health have nothing to do with the topic of the dialogue, even though these statements can be expressed in the same language with which statements about Donald Duck and his capability to think rationally are expressed. Thus, we would like to have, for any particular dialogue, a set of facts that have to do with the topic of the dialogue. Such a set is called a *knowledge base*.

**Definition 3.3.2** [Knowledge bases]. A *knowledge base* in an  $AS = (L_t, R, n)$  is a set  $\mathcal{K} \subseteq L_t$  consisting of two disjoint subsets  $\mathcal{K}_n$  (the *axioms*) and  $\mathcal{K}_p$  (the *ordinary premises*).

In the next definition, the notions of an argumentation system and a knowledge base are combined with that of an ordering on arguments. Below  $\preceq$  is a partial preorder such that  $A \preceq B$  means that  $B$  is at least as ‘good’ as  $A$ .  $A \prec B$  means  $A \preceq B$  and  $B \not\preceq A$  here.

**Definition 3.3.3** [Argumentation theory]. An *argumentation theory* (AT) is a triple  $AT = (AS, \mathcal{K}, \preceq)$  where  $AS$  is an argumentation system,  $\mathcal{K}$  is a knowledge base in  $AS$  and  $\preceq$  is an argument ordering on the set of all arguments that can be constructed from  $\mathcal{K}$  in  $AS$ .

From the contents of a knowledge base, dialogue participants can construct *arguments*. These arguments can contain two types of inference rules: *strict* inference rules, which are of the form “ $\varphi$  always follows from  $\Phi$ ”, where  $\Phi$  is a set of formulas, and *defeasible* inference rules, which are of the form “ $\varphi$  usually follows from  $\Phi$ ”, where  $\Phi$  is, again, a set of formulas. Whereas an argument can be undercut on a defeasible inference rule used in it, it cannot be undercut on one of its strict inference rules. Also, arguments can contain *subarguments*. For example: the argument  $A = (\psi \Rightarrow \chi) \rightarrow \varphi$ , where ‘ $\rightarrow$ ’ is a strict inference rule and ‘ $\Rightarrow$ ’ a defeasible inference rule, can be described in terms of its subarguments by letting  $A = B \rightarrow \varphi$ ,  $B = C \Rightarrow \chi$  and  $C = \psi$ , where  $B$  and  $C$  are subarguments of  $A$ , and  $C$  is a subargument of  $B$ .

In what follows, for a given argument, the function *Prem* returns all the formulas of  $\mathcal{K}$  (called *premises*) used to build the argument, *Conc* returns its conclusion, *Sub* returns all its sub-arguments, *DefRules* returns all the defeasible rules used in the argument and *TopRule* returns the last inference rule used in the argument.

**Definition 3.3.4** [Argument]. An *argument*  $A$  on the basis of an argumentation theory with a knowledge base  $\mathcal{K}$  and an argumentation system  $(L_t, R, n)$  is:

1.  $\varphi$  if  $\varphi \in \mathcal{K}$  with:  $Prem(A) = \{\varphi\}$ ,  $Conc(A) = \varphi$ ,  $Sub(A) = \{\varphi\}$ ,  $DefRules(A) = \emptyset$ ,  $TopRule(A) = \text{undefined}$ .
2.  $A_1, \dots, A_n \rightarrow \psi$  if  $A_1, \dots, A_n$  are arguments such that there exists a strict rule  $Conc(A_1), \dots, Conc(A_n) \rightarrow \psi$  in  $\mathcal{R}_s$ .  
 $Prem(A) = Prem(A_1) \cup \dots \cup Prem(A_n)$ ,  
 $Conc(A) = \psi$ ,  
 $Sub(A) = Sub(A_1) \cup \dots \cup Sub(A_n) \cup \{A\}$ ,  
 $DefRules(A) = DefRules(A_1) \cup \dots \cup DefRules(A_n)$ ,  
 $TopRule(A) = Conc(A_1), \dots, Conc(A_n) \rightarrow \psi$ .

3.  $A_1, \dots, A_n \Rightarrow \psi$  if  $A_1, \dots, A_n$  are arguments such that there exists a defeasible rule  $Conc(A_1), \dots, Conc(A_n) \Rightarrow \psi$  in  $\mathcal{R}_d$ .
- $$Prem(A) = Prem(A_1) \cup \dots \cup Prem(A_n),$$
- $$Conc(A) = \psi,$$
- $$Sub(A) = Sub(A_1) \cup \dots \cup Sub(A_n) \cup \{A\},$$
- $$DefRules(A) = DefRules(A_1) \cup \dots \cup DefRules(A_n) \cup \{Conc(A_1), \dots, Conc(A_n) \Rightarrow \psi\},$$
- $$TopRule(A) = Conc(A_1), \dots, Conc(A_n) \Rightarrow \psi.$$

**Examples:**

- $\varphi$  is a simple argument, which does not use any inference rules. In this argument, the premise is equal to the conclusion.
- $\varphi, \varphi \supset \psi \rightarrow_{mp} \psi$  is an argument in which a strict inference rule from propositional logic, called *modus ponens* ( $\rightarrow_{mp}$ ), is used to infer  $\psi$  from  $\varphi$  and  $\varphi \supset \psi$ . The strict inference expresses that  $\psi$  *always* follows from  $\varphi$  and  $\varphi \supset \psi$  through modus ponens.
- $\varphi \Rightarrow \psi$  is an argument in which the defeasible inference rule ( $\Rightarrow$ ) is used. The defeasible inference expresses that  $\psi$  *usually* follows from  $\varphi$  through ( $\Rightarrow$ ).

Sometimes an argument can contain premises which are not actually used in the inference that is expressed. For example, the inference  $\chi, \varphi, \varphi \supset \psi \rightarrow_{mp} \psi$  is valid in propositional logic, but the premise  $\chi$  is not necessary for the inference: it could just as well be left out. In [6], Diana Grooters defines a special type of argument, called a *minimal argument*, which takes only premises that are relevant for the inference.

**Definition 3.3.5** [Minimal argument]. A *minimal argument*  $A$  on the basis of an argumentation theory with a knowledge base  $\mathcal{K}$  and an argumentation system  $(L_t, R, n)$  is:

1.  $\varphi$  if  $\varphi \in \mathcal{K}$  with:  $Prem(A) = \{\varphi\}$ ,  $Conc(A) = \varphi$ ,  $Sub(A) = \{\varphi\}$ ,  $DefRules(A) = \emptyset$ ,  $TopRule(A) = \text{undefined}$ .
2.  $A_1, \dots, A_n \rightarrow \psi$  if  $A_1, \dots, A_n$  are minimal arguments such that there exists a strict rule  $Conc(A_1), \dots, Conc(A_n) \rightarrow \psi$  in  $\mathcal{R}_s$  and there is not a strict rule  $a_1, \dots, a_i \rightarrow \psi$  for  $\{a_1, \dots, a_i\} \subset Conc(\{A_1, \dots, A_n\})$ .  
 $Prem(A) = Prem(A_1) \cup \dots \cup Prem(A_n)$ ,  
 $Conc(A) = \psi$ ,  
 $Sub(A) = Sub(A_1) \cup \dots \cup Sub(A_n) \cup \{A\}$ ,  
 $DefRules(A) = DefRules(A_1) \cup \dots \cup DefRules(A_n)$ ,  
 $TopRule(A) = Conc(A_1), \dots, Conc(A_n) \rightarrow \psi$ .

3.  $A_1, \dots, A_n \Rightarrow \psi$  if  $A_1, \dots, A_n$  are minimal arguments such that there exists a defeasible rule  $Conc(A_1), \dots, Conc(A_n) \Rightarrow \psi$  in  $\mathcal{R}_d$ .
- $$Prem(A) = Prem(A_1) \cup \dots \cup Prem(A_n),$$
- $$Conc(A) = \psi,$$
- $$Sub(A) = Sub(A_1) \cup \dots \cup Sub(A_n) \cup \{A\},$$
- $$DefRules(A) = DefRules(A_1) \cup \dots \cup DefRules(A_n) \cup \{Conc(A_1), \dots, Conc(A_n) \Rightarrow \psi\},$$
- $$TopRule(A) = Conc(A_1), \dots, Conc(A_n) \Rightarrow \psi.$$

**Example:** assuming that  $R_s$  contains all inference rules which are valid in the classical interpretation of propositional logic,  $A = \psi, \psi \supset \varphi, \varphi \rightarrow \varphi$  is a non-minimal argument, since  $\psi, \psi \supset \varphi, \varphi \rightarrow \varphi \in R_s$ , but also  $\psi, \psi \supset \varphi \rightarrow \varphi \in R_s$  and  $\varphi \rightarrow \varphi \in R_s$ , for both of which it holds that it is a strict inference rule with a set of premises which is a strict subset of the set of conclusions of the subarguments of  $A$ .

**Definition 3.3.6** [Argument properties]. An argument  $A$  is *strict* if  $DefRules(A) = \emptyset$ ; *defeasible* if  $DefRules(A) \neq \emptyset$ ; *firm* if  $Prem(A) \subseteq \mathcal{K}_n$ ; *plausible* if  $Prem(A) \cap \mathcal{K}_p \neq \emptyset$ .

Next, notions of attack, succesful attack and defeat are defined. These notions can be seen as refinements of the concept of attack used in Dung-style argumentation frameworks, since they specifies not only that an argument attacks another argument, but also respectively how it attacks that argument, under what conditions such an attack is succesful and what it means for an argument to be defeated.

**Definition 3.3.7** [Attacks].  $A$  *attacks*  $B$  iff  $A$  *undercuts*, *rebuts* or *undermines*  $B$ , where:

- $A$  *undercuts* argument  $B$  (on  $B'$ ) iff  $Conc(A) = -n(r)$  for some  $B' \in Sub(B)$  such that  $B'$ 's top rule  $r$  is defeasible.
- $A$  *rebuts* argument  $B$  (on  $B'$ ) iff  $Conc(A) = -\varphi$  for some  $B' \in Sub(B)$  of the form  $B'_1, \dots, B'_n \Rightarrow \varphi$ .
- Argument  $A$  *undermines*  $B$  (on  $\varphi$ ) iff  $Conc(A) = -\varphi$  for an ordinary premise  $\varphi$  of  $B$ .

**Definition 3.3.8** [Successful rebuttal, successful undermining and defeat]

- $A$  *successfully rebuts*  $B$  if  $A$  rebuts  $B$  on  $B'$  and  $A \not\prec B'$ .
- $A$  *successfully undermines*  $B$  if  $A$  undermines  $B$  on  $\varphi$  and  $A \not\prec B$ .

- $A$  *defeats*  $B$  iff  $A$  undercuts, successfully rebuts or successfully undermines  $B$ .

Finally, argumentation theories can be linked to Dung-style argumentation frameworks.

**Definition 3.3.9** [Argumentation framework] An *abstract argumentation framework* (AF) corresponding to an argumentation theory AT is a pair  $\langle \mathcal{A}, Def \rangle$  such that:

- $\mathcal{A}$  is the set of arguments on the basis of  $AT$  as defined by Definition 3.3.4.
- $Def$  is the relation on  $\mathcal{A}$  given by Definition 3.3.8.

From a theoretical viewpoint, the correspondence between an AT and an AF does not appear complex at all. However, the situation changes when we move to the practical viewpoint. In a practical setting, one gets from an AT to an AF by converting the first into the second by means of some algorithm. Now, consider the case in which we have a singleton knowledge base  $\mathcal{K} = \{\varphi\}$  and a singleton set of strict inference rules  $R_s = \{r : \varphi \rightarrow \varphi \cup \psi\}$ . Although rule  $r$  is a valid inference rule in propositional logic, it poses a big challenge for an algorithm designed to convert an AT to an AF, since for this conversion, the set  $S$  of all arguments which can be constructed on the basis of the AT needs to be obtained. Let us try to get  $S$  for our example. We start with  $S = \{A_1 : \varphi\}$ , which contains only one simple argument  $A_1$  that has the proposition  $\varphi$  as both its premise and its conclusion. Now, we can generate a new argument  $(A_2 : \varphi \cup \psi \text{ since } A_1, r)$  from  $A_1$ , since  $\varphi \cup \psi$  can be inferred from feeding  $\varphi$  as input to  $\varphi \rightarrow \varphi \cup \psi$ . Thus, our new  $S$  is equal to  $\{A_1 : \varphi, A_2 : \varphi \cup \psi \text{ since } A_1, r\}$ . Next, we can do the same trick again for  $A_2$ , which gives us  $S = \{A_1 : \varphi, A_2 : (\varphi \cup \psi \text{ since } A_1, r), A_3 : (\varphi \cup \psi) \cup \psi \text{ since } A_2, r\}$ , and for  $A_3$ , and for  $A_4$ , and so on forever. The conclusion is that this algorithm will never terminate, which means that we never get from our AT to its corresponding AF.

We can solve the problem of an endlessly running algorithm by putting restrictions on our inference rules. For example: we could allow only inference rules of which the conclusion contains one operator less than the premise of the rule with the highest number of operators of all premises of the rule. This way, the algorithm will, under the assumption of course that  $\mathcal{K}$  and  $R_s$  are finite, never run forever. However, further investigation of such restrictions on inference rules is outside the scope of this research. Therefore, we just assume here that we have an algorithm called “convertATtoAF” that

converts an AT to an AF in the way that is described in 3.3.9, which has a big O notation complexity of  $C(\text{convertATtoAF})$ .

Later on in the thesis, complexity values of other algorithms will be described in terms of  $C(\text{convertATtoAF})$  and variables indicating the size of certain data structures which are fed as input to those algorithms. Sometimes two complexity values  $C(\text{str}_1) = O(\alpha)$  and  $C(\text{str}_2) = O(\beta)$ , where  $\text{str}_1$  and  $\text{str}_2$  are names of algorithms and  $\alpha$  and  $\beta$  are numerical expressions which may contain variables, need to be added up. In that case, we write  $C(\text{str}_1) + C(\text{str}_2)$ , which yields a complexity value of  $O(\alpha + \beta)$ . Also,  $n \times C(\text{str}_1)$  is written sometimes, which yields  $O(n \times \alpha)$ , and  $n + C(\text{str}_1)$ , yielding  $O(n + \alpha)$ .

It is now possible to define a consequence notion for well-formed formulas.

**Definition 3.3.10** [Acceptability of conclusions] For any semantics  $S$  and for any argumentation theory  $AT$  and formula  $\varphi \in \mathcal{L}_{AT}$ :

1.  $\varphi$  is *sceptically S-acceptable* in  $AT$  if and only if there exists an argument with conclusion  $\varphi$  that is contained in all  $S$ -extensions of  $AT$ .
2.  $\varphi$  is *credulously S-acceptable* in  $AT$  if and only if there exists an  $S$ -extension of  $AT$  that contains an argument with conclusion  $\varphi$ .

We can imagine an algorithm for determining whether some proposition is sceptically or credulously  $S$ -acceptable in some AT for some semantics  $S$ , for which we need to be able to generate all extensions of that AT. Just like in the case of converting an AT to an AF, it is beyond the scope of this thesis to actually present such an algorithm, but again we give the algorithm a name, “generateExtensionsFromAT”, and we say that the big O notation complexity of the algorithm is  $C(\text{generateExtensionsFromAT})$ . Next, we assume we have algorithms “isScepticallyAcceptable” and “isCredulouslyAcceptable” for determining whether some proposition is respectively sceptically or credulously  $S$ -acceptable in some AT under some semantics  $S$ . The complexity  $C(\text{testAcceptability})$  of both these algorithms can be obtained by simply adding up  $C(\text{convertATtoAF})$ ,  $C(\text{generateExtensionsFromAT})$  and  $O(n)$ , where  $n$  is the number of extensions of the AT which is fed to the algorithms as input. This is the complexity, since both algorithms first convert their input AT to an AF, generate all extensions of the AF after that, and then walk through all the extensions once whilst performing the actual acceptability tests.

The following corollary derived from Definition 3.3.10, which will be referred to later on in the thesis, is one developed in this thesis:

**Corollary 3.3.11** If argumentation theory  $AT$  has only one extension, then it holds for any formula  $\varphi$  that, given a semantics  $S$  under which Dung-style argumentation frameworks are interpreted,  $\varphi$  is credulously  $S$ -acceptable in  $AT$  if and only if  $\varphi$  is sceptically  $S$ -acceptable in  $AT$ .

**Proof:**

The proof for this corollary is straightforward. When, given semantics  $S$ , a formula is credulously  $S$ -acceptable in the abstract argumentation theory  $AT$ , this means that the formula is in some extension of  $AT$ . If  $AT$  has only one extension, this automatically means that the formula is in all extensions of  $AT$ , which implies that the formula is sceptically  $S$ -acceptable in  $AT$ . Similarly, when a formula is sceptically  $S$ -acceptable in argumentation theory  $AT$ , this means that the formula is in all extensions of  $AT$ , which automatically means that the formula is in some extension of  $AT$ , which implies that the formula is credulously  $S$ -acceptable in  $AT$ . This concludes the proof.  $\square$

Corollary 3.3.11 is a strong result, since it implies that, in the case of an argumentation theory that has only one extension, the difference between credulous and sceptical acceptability disappears. As it turns out, there is a large and important class of argumentation theories which always have one extension. This is the class of AT's from which the set of arguments  $Args$  can be constructed, for which there is no infinite sequence  $A_1, \dots, A_n, \dots$  such that for each  $i$ ,  $A_i \in Args$  and  $A_{i+1}$  attacks  $A_i$ . In [3], the proof for Theorem 30 shows that each member of the class of argumentation frameworks corresponding to this class, which Dung calls the class of *well-founded* AF's, has exactly one complete extension which is also grounded, stable and preferred.

Let us look at an example where there is, in fact, a difference between the notions of sceptical and credulous acceptability, and see if this situation requires an argumentation theory that has multiple extensions. Suppose we have the following abstract argumentation framework  $AF$  which is based on some argumentation theory  $AT$ :

$$A \longleftrightarrow B$$

As was already stated in Section 3.2, where the same example was given,  $AF$  has two extensions under preferred semantics:  $\{A\}$  and  $\{B\}$ . Clearly, the notions of sceptical and credulous acceptability behave differently here, since  $Conc(A)$  and  $Conc(B)$  are both in some extension of  $AF$ , but not in

all of them.

### 3.4 Prakken’s framework formally defined

This section contains the definitions regarding Prakken’s framework as they are presented in [8]. Most of the concepts which are defined here were already discussed informally in Section 2.2.

We start with a dialogue topic  $t \in L_t$  (where  $L_t$  is the topic language), a *proponent* ( $P$ ) who defends  $t$  and an *opponent* ( $O$ ) who challenges  $t$ . For any player  $p$ , we define:

- $\bar{p} = O$  iff  $p = P$ , and
- $\bar{p} = P$  iff  $p = O$ .

The following definition is the top definition of a dialogue system:

**Definition 3.4.1** [Dialogue games for argumentation]. A *dialogue system for argumentation* (dialogue system for short) is a pair  $(\mathcal{AS}, \mathcal{D})$ , where  $\mathcal{AS}$  is an  $ASPIC^+$  argumentation system and  $\mathcal{D}$  is a dialogue system proper.

From here, we move on to the other elements of dialogue systems.

**Definition 3.4.2** A *dialogue system proper* is a triple  $\mathcal{D} = (L_c, P, C)$  where  $L_c$  (the communication language) is a set of locutions,  $P$  is a *protocol* for  $L_c$ , and  $C$  is a set of effect rules of locutions in  $L_c$ , specifying the effects of the locutions on the participants’ *commitments*.

**Definition 3.4.3** A *communication language* is a tuple  $L_c = (S, R_a, R_s)$ , where  $S$  is a set of locutions and  $R_a$  and  $R_s$  are two binary relations of *attacking* and *surrendering reply* on  $S$ . Each  $s \in S$  is of the form  $p(c)$  where  $p$  is an element of a given set  $P$  of performatives and  $c$  either is a member or subset of  $L_t$ , or is a member of  $Args$  (of some given logic  $L$ ). Both  $R_a$  and  $R_s$  are irreflexive (no locution can be used to attack or surrender to itself) and in addition satisfy the following conditions:

1.  $R_a \cap R_s = \emptyset$  (no locution both attacks and surrenders to another one);
2.  $\forall a, b, c : (a, b) \in R_a \Rightarrow (a, c) \notin R_s$  (a locution that attacks some other locution cannot also be used to surrender to some other locution);
3.  $\forall a, b, c : (a, b) \in R_s \Rightarrow (c, a) \notin R_a$  (a locution that surrenders to some other locution cannot be attacked any more by some other locution).

The function  $att : R_s \rightarrow \mathcal{P}(R_a)$  assigns to each pair  $(a, b) \in R_s$  one or more *attacking counterparts*  $(c, b) \in R_a$ .

**Definition 3.4.4** [Moves and dialogues].

- The set  $M$  of *moves* is defined as  $\mathbb{N} \times \{P, O\} \times L_c^p \times \mathbb{N}$ , where the four elements of a move  $m$  are denoted by, respectively:
  - $id(m)$ , the *identifier* of the move,
  - $pl(m)$ , the *player* of the move,
  - $s(m)$ , the *speech act* performed in the move,
  - $t(m)$ , the *target* of the move.

For example: the move  $m = (2, P, argue(A), 1)$  is placed by player  $P$ , has identifier 2, targets the move with identifier 1 and has as its content the speech act  $argue(A)$ .

- The set of *dialogues*, denoted by  $M^{\leq \infty}$ , is the set of all sequences  $m_1, \dots, m_i, \dots$  from  $M$  such that
  - each  $i^{th}$  element in the sequence has identifier  $i$ ;
  - $t(m_1) = 0$ ;
  - for all  $i > 1$  it holds that  $t(m_i) = j$  for some  $m_j$  preceding  $m_i$  in the sequence.

The set of *finite dialogues*, denoted by  $M^{< \infty}$ , is the set of all finite sequences that satisfy these conditions. For any dialogue  $d = m_1, \dots, m_n, \dots$ , the sequence  $m_1, \dots, m_i$  is denoted by  $d_i$ , where  $d_0$  denotes the empty dialogue.

When  $t(m) = id(m')$ , Prakken says in [8] that  $m$  *replies to*  $m'$  in  $d$  and also that  $m'$  is the *target of*  $m$  in  $d$ . Also, he sometimes slightly abuses notation and lets  $t(m)$  denote a move instead of just its identifier. When  $s(m)$  is an attacking (surrendering) reply to  $s(m')$ , he also says that  $m$  is an attacking (surrendering) reply to  $m'$ .

**Definition 3.4.5** A *turntaking function*  $T$  is a function

- $T : M^{< \infty} \rightarrow \mathcal{P}(\{P, O\})$ .

such that  $T(\emptyset) = \{P\}$  (the proponent always begins). A *turn* of a dialogue is a maximal sequence of stages in the dialogue where the same player moves.

When  $T(d)$  is a singleton, the brackets will be omitted.

**Definition 3.4.6** A *protocol* on  $M$  is a function  $Pr$  with domain a nonempty subset of  $M^{<\infty}$  taking subsets of  $M$  as values. The elements of  $dom(Pr)$  (the domain of  $Pr$ ) are called the *legal finite dialogues*. The elements of  $Pr(d)$  are called the *moves allowed after  $d$* . If  $d$  is a legal dialogue and  $Pr(d) = \emptyset$ , then  $d$  is said to be a *terminated dialogue*.  $Pr$  must satisfy the following condition: for all finite dialogues  $d$  and moves  $m$  it holds that  $d \in dom(Pr)$  and  $m \in Pr(d)$  iff  $d, m \in dom(Pr)$ .

All protocols are further assumed to satisfy the following basic conditions for all moves  $m_i$  and all legal finite dialogues  $d$ .

If  $m \in Pr(d)$ , then:

- $R_1$  :  $pl(m) \in T(d)$ ;
- $R_2$ : If  $d \neq d_0$  and  $m \neq m_1$ , then  $s(m)$  is a reply to  $s(t(m))$  according to  $L_c$ ;
- $R_3$ : If  $m$  replies to  $m'$ , then  $pl(m) \neq pl(m')$ ;
- $R_4$ : If there is an  $m'$  in  $d$  such that  $t(m) = t(m')$  then  $s(m) \neq s(m')$ ;
- $R_5$ : For any  $m'$  in  $d$  that surrenders to  $t(m)$ ,  $m$  is not an attacking counterpart of  $m'$ .

Together these conditions capture a lower bound on coherence of dialogues. Rule  $R_1$  says that a move is legal only if moved by the player-to-move.  $R_2$  says that a replying move must be a reply to its target according to  $L_c$ , and  $R_3$  says that one cannot reply to one's own moves. Rule  $R_4$  states that if a player backtracks, the new move must be different from the first one. ('backtracking' in this article is taken to mean any alternative reply to the same target in a later turn). Finally,  $R_5$  says that surrenders may not be 'revoked'.

**Definition 3.4.7** A *commitment function* is a function

- $C : M^{\leq\infty} \times \{P, O\} \longrightarrow \mathcal{P}(L_t)$ .

such that  $C_\emptyset(p) = \emptyset$  (meaning that agents start a dialogue with no commitments).  $C_d(p)$  denotes player  $p$ 's commitments in dialogue  $d$ .

### 3.5 Liberal dialogue systems

This section contains the definitions regarding liberal dialogue systems as they are presented in [8].

A class of liberal dialogue systems is defined (parameterized by a logic  $\mathcal{L}$ ), in which the participants have much freedom, and which is intended to be the core of all other dialogue systems of Prakken’s study. The following holds for liberal dialogue systems:

- They have context-independent protocols.
- They model only pure persuasion.
- Their protocols have public semantics.
- They have multiple-move protocols.

Liberal dialogues greatly rely for their coherence on the cooperativeness of the dialogue participants.

The following table contains an overview of the speech acts that are available in liberal dialogue systems, and how they are linked to each other by relations of attacking and surrendering.

Acts	Attacks	Surrenders
claim $\varphi$	why $\varphi$	concede $\varphi$
why $\varphi$	argue A ( $\text{Conc}(A) = \varphi$ )	retract $\varphi$
argue A	why $\varphi$ ( $\varphi \in \text{Prem}(A)$ ) argue B (B defeats A)	concede $\varphi$ ( $\varphi \in \text{Prem}(A)$ or $\varphi = \text{Conc}(A)$ )
concede $\varphi$		
retract $\varphi$		

The following commitment rules apply to the speech acts in the communication language (below  $s$  denotes the speaker of the move).

- If  $s(m) = \text{claim}(\varphi)$  then  $C_s(d, m) = C_s(d) \cup \{\varphi\}$ ;
- If  $s(m) = \text{why}(\varphi)$  then  $C_s(d, m) = C_s(d)$ ;
- If  $s(m) = \text{concede}(\varphi)$  then  $C_s(d, m) = C_s(d) \cup \{\varphi\}$ ;
- If  $s(m) = \text{retract}(\varphi)$  then  $C_s(d, m) = C_s(d) - \{\varphi\}$ ;
- If  $s(m) = \text{argue}(A)$  then  $C_s(d, m) = C_s(d) \cup \text{Prem}(A) \cup \{\text{Conc}(A)\}$ .

Also, the following (very liberal) turntaking rule is adopted in liberal dialogue systems.

- $T_L : T(d_0) = P, T(d_1) = O, \text{ else } T(d) = \{P, O\}$ .

This means that the players are fixed only for the first and second moves ( $P$  must do the first move, while  $O$  must do the second). The rest of the moves can be done by either one of the players.

The protocol for liberal dialogue systems adds two more rules to the set of rules that was assumed in the previous section.

If  $m \in P(d)$ , then:

- $R_6$ : If  $d = \emptyset$ , then  $s(m)$  is of the form *claim*( $\varphi$ ) or *argue*( $A$ ).
- $R_7$ : If  $m$  concedes the conclusion of an argument moved in  $m'$ , then  $m'$  does not reply to a *why* move.

$R_6$  says that each dialogue begins with either a claim or an argument. The initial claim or, if a dialogue starts with an argument, its conclusion is the topic of the dialogue.  $R_7$  restricts concessions of an argument's conclusion to conclusions of counterarguments. This ensures that propositions are conceded at the place in which they were introduced.

**Definition 3.5.1** A *dialogue system for liberal dialogues* is now defined as any dialogue system with  $L_c$  as specified in Section 3.5, with turntaking rule  $T_L$  and such that a move is legal if and only if it satisfies protocol rules  $R_1$ - $R_7$ .

A dialogue is defined as *terminated* just in case no legal continuation is possible. However, realistic dialogues will often terminate earlier, by external agreement or decision to terminate it.

Since players can decide to end a dialogue at any time during the dialogue, 'any time' *outcome* definitions are needed.

**Definition 3.5.2** [Dialogical status of moves] All attacking moves in a finite dialogue  $d$  are either *in* or *out* in  $d$ . Such a move  $m$  is *in* iff

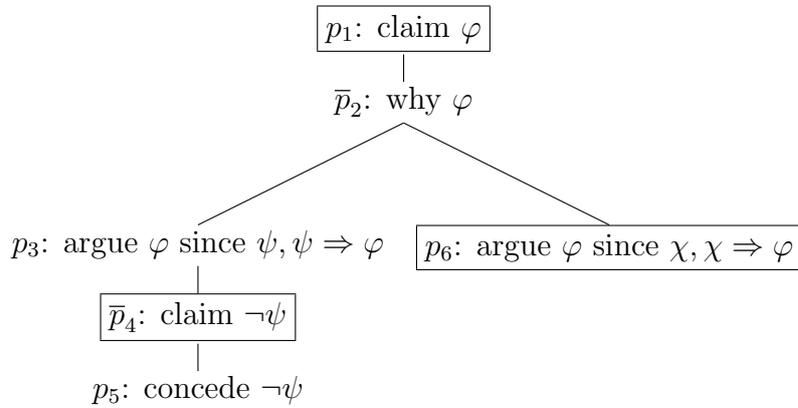
1.  $m$  is surrendered in  $d$ ; or else
2. all attacking replies to  $m$  are *out*

Otherwise  $m$  is *out*

**Definition 3.5.3** A move  $m$  in a dialogue  $d$  is *surrendered* in  $d$  iff

- it is an *argue*  $A$  move and it has a reply in  $d$  that concedes  $A$ 's conclusion; or else
- $m$  has a surrendering reply in  $d$ .

In the rest of this thesis, whenever an example dialogue is presented as a tree, moves that are *in* will be contained inside a box, whereas moves that are *out* will not. The following is an example of such a dialogue tree.



**Definition 3.5.4** [The current winner of a dialogue] The status of the initial move  $m_1$  of a dialogue  $d$  is *in favour of*  $P(O)$  and *against*  $O(P)$  iff  $m_1$  is *in* (*out*) in  $d$ . We also say that  $m_1$  favours, or is against  $p$ . Player  $p$  *currently wins* dialogue  $d$  if  $m_1$  of  $d$  favours  $p$ .

**Definition 3.5.5** For any dialogue  $d$  the proponent wins  $d$  if  $m_1$  is *in*, otherwise the opponent wins  $d$ .

In the example above Definition 3.5.4, the proponent wins the dialogue, since the initial move  $p_1$  of the dialogue by that proponent is *in*.

## 4 Desirable properties of dialogue agents

In this section, the first research question of this thesis is answered: “*how should participants of persuasion dialogues behave ideally?*”. We will use the pragmatic theory of Grice to define ideal behaviour of dialogue participants. Before we do this, however, we take a closer look at dialogue participants.

## 4.1 Dialogue participants

Participants in argumentation based persuasion dialogues are *agents*, which have a *belief base* that is defined as follows:

**Definition 4.1.1** Given a liberal dialogue system  $DS = (AS, D)$  with  $AS = (L_t, R, n)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$  and a dialogue  $d$  of  $DS$ , the *belief base* of agent  $p$  of  $d$  is defined as  $\mathcal{K}_p(p)$ , where  $\mathcal{K}_p(p) \subseteq \mathcal{K}_p \subseteq K$ .

In the rest of this thesis, an agent’s belief base is treated as a *private* object, in the sense that neither the other agent participating in  $d$  nor any rule in  $DS$ ’s protocol has access to it. Furthermore,  $R_s$ ,  $R_d$ ,  $\mathcal{K}_n$  and  $\preceq$  are treated as *public* objects, in the sense that both the agents participating in  $d$  and all protocol rules in  $DS$  have access to them.

## 4.2 Grice’s maxims

Ideal agents adhere to a number of dialogical conventions during a dialogue, which we identify in this section. As it happens, there exists a subbranch of philosophy of language which deals with such conventions, viz. pragmatics, which can help us identify dialogical conventions for persuasion dialogues. One of the historically most important works on pragmatics, is Grice’s paper on logic and conversation (see [5]), in which he defines the so-called *Cooperation Principle* and a set of conversational conventions or *maxims* (which are commonly called ‘Grice’s maxims of conversation’).

In [5], Grice first introduces a notion which he calls ‘conversational implicature’. The notion refers to the property of sentences uttered in a conversation to possibly imply multiple meanings in the context of that conversation. These meanings can be literal, but they can also be figurative. For example: if person  $A$  asks person  $B$  “what is the capital of the Netherlands?” and person  $B$  answers “the capital of the Netherlands is Amsterdam”, then  $B$ ’s answer has a literal meaning in which  $B$  actually tells  $A$  that Amsterdam is the capital of the Netherlands. On the other hand, if  $B$  answers “the capital of the Netherlands is ‘you’ve got Wikipedia for these sorts of questions’”, then, of course,  $B$  does not literally mean that the capital of the Netherlands is ‘you’ve got Wikipedia for these sorts of questions’; rather,  $B$  means to encourage  $A$  to look for the answer herself. According to Grice, which of the possible meanings of a sentence is the actual meaning conveyed by the speaker to the hearer in a conversation, depends on whether certain conversational conventions known to both participants of the conversation are violated or not. Grice calls these conventions ‘maxims of conversation’,

and he defines one supermaxim (the Cooperation Principle) and four classes of general maxims.

An interesting question is: can Grice’s maxims be used a guidelines for coherence in persuasion dialogues? In order for us to be able to answer this question, we need to look at the similarities and differences between the class of conversations to which Grice’s maxims apply and persuasion dialogues. In [5], Grice refers to the type of communication used in this class of conversations with the term “natural speech”. Although he does not explicitly mention the goal of this type of speech, information exchange appears to be the best candidate. Let us look at the similarities between what we will call “natural conversations” from here on, in which Grice’s natural speech is used, and persuasion dialogues. One clear similarity is that in both natural conversations and persuasion dialogues information is exchanged by means of utterances. The rules governing this exchange are, however, not the same for both types of communication. Bound to the goal of persuasion is the fact that participants of a persuasion dialogue strive to win the dialogue, whereas participants of natural conversations may not compete at all. Taking all this into account, we decide here that we can use Grice’s maxims to capture conventions for persuasion dialogues which involve aspects of information exchange, but that we also need other conventions involving the aspect of winning that is characteristic for persuasion.

Grice’s maxims are introduced below. For each of the maxims, it is explained why it can or cannot be applied to the domain of persuasion and if it can be applied, how this can be done.

**The Cooperation Principle** Grice defines the Cooperation Principle as follows: “*make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged*”. This supermaxim is adhered to by an individual if that individual adheres to all other maxims. In the same way, the maxim will be adhered to in persuasion dialogues by an agent if that agent adheres to the other maxims in those dialogues.

**The maxims of quantity** Grice defines two maxims of this kind:

- “*Make your contribution as informative as is required (for the current purposes of the exchange)*”. It is difficult to apply this maxim to Prakken’s framework, since there is no notion of (quantity of) information in the framework.
- “*Do not make your contribution more informative than is required*”. The same holds here as in the case of the first maxim of quantity.

**The maxims of quality** He also defines two maxims of quality:

- “*Do not say what you believe to be false*”. In persuasion dialogues, this means that agents should not commit to some proposition, while actually they believe its negation, or say that they are ignorant about some proposition, while in fact they are not.
- “*Do not say that for which you lack adequate evidence*”. In persuasion dialogues, this means that agents should not commit to some proposition, while they actually have no beliefs about it.

**The maxims of relation** Grice places a single maxim under this category: “*be relevant*”. In persuasion dialogues, this means that an agent’s every move should contribute somehow to that agent’s goal of winning the dialogue.

**The maxims of manner** Here, Grice gives “*be perspicuous*” as a super-maxim. Furthermore, he gives the following other maxims:

- “*Avoid obscurity of expression*”. This maxim has no application in Prakken’s framework.
- “*Avoid ambiguity*”. This maxim has no application in Prakken’s framework.
- “*Be brief (avoid unnecessary prolixity)*”. In persuasion dialogues, this means that agents’ moves should contain no unnecessary information.
- “*Be orderly*”. This maxim has no application in Prakken’s framework.

Note that, after he has presented the maxims of manner in [5], he says: “*and one might need others*”.

Next, we wish to create a maxim of our own for the aspect of winning in the domain of persuasion. Here, two important things should be mentioned. The first is that this new maxim will be different from the other maxims, since the other maxims are all about aspects of coherence of persuasion dialogues which are also found in Grice’s natural conversation, whereas the new maxim is about winning, which is not an aspect of natural conversation. The second thing is that two types of winning play a part in persuasion dialogues: a dialogue agent’s individual winning and the collective winning of both agents when an agreement is reached at the end of a dialogue. For these two types of winning, it holds that the former type should be in service of the latter,

meaning that agents should try to win individually, but always with the aim of reaching an agreement. Hence, a maxim about the winning aspects of persuasion dialogues should say that agents need to put as much effort as possible into trying to win a dialogue individually, but that they should achieve this by behaving properly, meaning that they should respect the other maxims for persuasion dialogues.

We end up with the following list of maxims for persuasion dialogue agents:

- Be honest.
  - Do not say what you believe to be false.
  - Do not say that for which you lack adequate evidence.
- Be relevant.
- Be as brief as possible (do not display “verbosity”).
- Within the bounds of proper behaviour, do your best to win individually (make “optimally useful” moves).

In the following three sections, the second research question of this thesis is answered: “*depending on the design choices made, what types of improper behaviour of dialogue participants can occur in Prakken’s framework?*”. While Section 5 treats liberal dialogues as defined in Section 3.5, Sections 6 and 7 deal with variants of liberal dialogues that have different communication languages.

## 5 Improper behaviour of agents in liberal dialogues

This section contains an overview of types of dishonest, irrelevant, verbose and non-optimally useful moves in liberal dialogue systems, under the assumptions we made so far about liberal dialogue systems and the agents participating in them. Definitions for several types of improper behaviour will be based on definitions of other notions, such as those of a commitment based argumentation theory, of a belief based argumentation theory and of persuasiveness.

## 5.1 Preliminary definitions

Two variants of the concept of an argumentation theory, which was captured in Definition 3.3.3, are defined here. Whereas the first of these variants is based on the commitments of an agent, the second is based on an agent's beliefs.

**Definition 5.1.1** Given a liberal dialogue system  $DS = (AS, D)$  with  $AS = (L_t, R, n)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$  and a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$ , the *commitment based argumentation theory of participant  $p$*  is:  $CAT_d(p) = (AS, C_d(p), \preceq)$ .

**Definition 5.1.2** Given a liberal dialogue system  $DS = (AS, D)$  with  $AS = (L_t, R, n)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$  and a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$ , the *belief based argumentation theory of participant  $p$*  is:  $BAT(p) = (AS, \mathcal{K}_p(p), \preceq)$ , where  $\mathcal{K}_p(p)$  are the individual beliefs of  $p$ .

The difference between a commitment based and a belief based argumentation theory is that the information contained in the former can always be accessed by protocol rules of a dialogue system respecting *public semantics*, while this does not hold for the latter since it involves agents' individual, local beliefs.

The big O notation complexity of an algorithm that generates either a commitment or a belief based argumentation theory is  $O(1)$ , since all the algorithm needs to do is combine parts of its input into a new data structure.

From this, we move on to persuasiveness and the definitions of two concepts involved in it.

**Definition 5.1.3** An *attackable element* of an argument  $A$  is either:

- the conclusion of  $A$  if  $A$  is of the form  $A'_1, \dots, A'_n \Rightarrow \varphi$ , or
- a premise  $\psi$  of  $A$  if  $\psi$  is an ordinary premise, or
- a defeasible inference rule  $r$  used in  $A$ .

Assuming that it can be determined in  $O(1)$  time that a premise is ordinary and that a rule is defeasible, one could create an algorithm for checking the attackability of an element that runs in  $O(1)$  time, since the algorithm has three cases, two of which involve either determination that a premise is ordinary or that a rule is defeasible, and a third that involves determining whether an argument has a particular form, which can also be done in  $O(1)$ .

**Definition 5.1.4** Given a liberal dialogue system  $DS = (AS, D)$  with  $AS = (L_t, R, n)$ , a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$ , a commitment based argumentation theory  $CAT_d(p)$ , semantics  $S$  for interpreting Dung-style argumentation frameworks and an argument  $A$  by  $\bar{p}$  which is a legal continuation of  $d$ , an *obviously sceptically persuasive element* of  $A$  is either:

- the conclusion of  $A$  if  $A$  is of the form  $A'_1, \dots, A'_n \Rightarrow \varphi$  and  $\varphi$  is sceptically  $S$ -acceptable in  $CAT_d(p)$ , or
- a premise  $\psi$  of  $A$  if  $\psi$  is an ordinary premise and  $\psi$  is sceptically  $S$ -acceptable in  $CAT_d(p)$ , or
- a defeasible inference rule  $r$  used in  $A$  if  $p$  can construct from  $CAT_d(p)$  an argument which is justified according to semantics  $S$  and in which  $r$  is used.

There is also the definition of an *obviously credulously persuasive element*, which is obtained by taking the definition of a sceptically persuasive element and replacing “sceptically” by “credulously” and “justified” by “defensible” everywhere in the definition. Moreover, both definitions have *unobvious* variants, which are obtained by replacing “ $CAT_d(p)$ ” by “ $BAT(p)$ ” and “obviously” by “unobviously” everywhere in the definitions.

One could create an algorithm for determining whether an attackable element is a persuasive element of some kind or not, with a complexity that is equal to  $C(\text{testAcceptability})$ . For the cases in which the input element is a conclusion or a premise, the complexity is equal to  $C(\text{testAcceptability})$  simply because in these cases the *testAcceptability* algorithm itself which is called here is the most complex step. When the input element is a defeasible inference rule, two of the three steps taken in the algorithm are the same as in the *testAcceptability* algorithm, i.e. converting an AT to an AF and generating all extensions of that AF. The third step is different, but it also involves visiting all extensions of the generated AF once, this time in order to determine which justified arguments exist and if one of these justified arguments contains the defeasible inference rule which was fed as input to the algorithm. Hence, the third step has a complexity of  $O(n)$ , where  $n$  is the number of extensions of the AF generated by the algorithm, which means that the complexity of the third case of our algorithm which implements Definition 5.1.4 and the overall complexity of the algorithm are equal to  $C(\text{testAcceptability})$ .

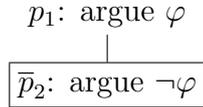
**Definition 5.1.5** Given a liberal dialogue system  $DS = (AS, D)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$ , a dialogue  $d$  of  $DS$  with participants  $p$  and

$\bar{p}$ , a commitment based argumentation theory  $CAT_d(p)$ , semantics  $S$  for interpreting Dung-style argumentation frameworks and a move  $m$  containing the speech act ‘argue  $A$ ’ by  $p$  which is a legal continuation of  $d$ , the *obvious sceptical persuasiveness* of  $m$  for  $\bar{p}$  is  $SPE/AE$  in the case where  $AE \neq 0$ , and 1 in the case where  $AE = 0$ , where  $SPE$  is the number of obviously sceptically persuasive elements in  $A$ , and  $AE$  is the number of attackable elements in  $A$ .

This definition also has a credulous variant, namely *obvious credulous persuasiveness*, which is obtained by replacing the word ‘sceptical’ by ‘credulous’, ‘sceptically’ by ‘credulously’ and ‘SPE’ by ‘CPE’ in the definition. Moreover, also these definitions both have a *unobvious* variant, which are obtained by replacing ‘ $CAT_d(p)$ ’ by ‘ $BAT(p)$ ’, ‘obviously’ by ‘unobviously’ and ‘obvious’ by ‘unobvious’ everywhere in the definitions.

An algorithm ‘calcPersuasiveness’ for calculating the persuasiveness of some kind of an argument is conceivable, which needs to determine for each of the argument’s attackable elements whether that element is persuasive in some way or not. Hence, if  $AE$  equals the number of attackable elements of the argument, then the complexity of the algorithm is equal to  $AE \times C(\text{testAcceptability})$ .

**Example:** We start with a liberal dialogue system  $D$  and the following dialogue  $d$ :



Now, what are, for  $p$ , the obvious and unobvious sceptical persuasiveness values of  $\bar{p}_2$ ’s argument? And what are the obvious and unobvious credulous persuasiveness values of this same argument for  $p$ ?

Before we can answer these questions, we first need to make some additional assumptions. We assume that  $K_p(p) = \{\varphi, \neg\varphi\}$  (an inconsistent belief base),  $K_p(\bar{p}) = \{\neg\varphi\}$  and  $R_d = R_s = \emptyset$ . Also, we take preferred semantics as the semantics  $S$  under which our Dung-style argumentation frameworks are interpreted. Last, we assume that both  $\varphi$  and  $\neg\varphi$  have the same preference.

Using Definition 3.3.9, we can now derive the argumentation framework  $AF_p$  from  $p$ ’s belief based argumentation theory  $BAT(p)$ :  $A_1 : \varphi$  and  $A_2 : \neg\varphi$  are the arguments in  $AF_p$ , and these arguments attack each other. Under preferred semantics, we obtain two possible extensions:  $\{A_1\}$  and  $\{A_2\}$ . We can also derive a second argumentation framework  $AF'_p$  from  $p$ ’s commitment based argumentation theory  $CAT_{p_1}(p)$ :  $A : \varphi$  is the only argument in  $AF'_p$

and it does not attack itself. Under preferred semantics (and also under the other types of semantics, since the attack relation of  $AF'_p$  is acyclic), we obtain one possible extension:  $\{A\}$ .

Now, we move on to the definitions that have to do with persuasiveness.  $\bar{p}_2$ 's argument contains one attackable element, namely  $\neg\varphi$ , which is both the premise and the conclusion of the argument. Hence, if we apply Definition 5.1.5 to our example, we get  $AE = 1$ . Since  $\neg\varphi$  is in one of the extensions of  $AF_p$ , it is credulously  $S$ -acceptable in  $CAT_{p_1}(p)$ . However,  $\neg\varphi$  is not sceptically  $S$ -acceptable in  $CAT_{p_1}(p)$ , since it is not in all extensions of  $AF_p$ . Thus,  $\bar{p}_2$ 's argument has one unobviously credulously persuasive element, but zero unobviously sceptically persuasive elements, which makes  $SPE = 0$  and  $CPE = 1$ .

We now have enough information to answer our initial questions: the unobvious sceptical persuasiveness of  $\bar{p}_2$ 's argument for  $p$  is  $SPE/AE = 0/1 = 0$ , while the unobvious credulous persuasiveness of  $\bar{p}_2$ 's argument for  $p$  is  $CPE/AE = 1/1 = 1$ . As for the obvious variants of the definitions: since  $\neg\varphi$  is in none of the extensions of  $AF'_p$ , it is neither credulously nor sceptically  $S$ -acceptable in  $CAT_{p_1}(p)$ . Thus,  $\bar{p}_2$ 's argument has zero obviously credulously persuasive elements and zero obviously sceptically persuasive elements, which makes  $SPE = 0$  and  $CPE = 0$ . Thus, the obvious sceptical persuasiveness of  $\bar{p}_2$ 's argument for  $p$  is  $SPE/AE = 0/1 = 0$  and the obvious credulous persuasiveness of  $\bar{p}_2$ 's argument for  $p$  is  $CPE/AE = 0/1 = 0$ .

So far, we have seen only two numeric values as results in our example: 0 and 1. One might wonder, now: what is the meaning of these values? What does it mean when an argument has a persuasiveness (for simplicity, we use this term here for all the obvious, unobvious, sceptical and credulous variants) of 0 or a persuasiveness of 1? And what about a persuasiveness value of 0.5? For the obvious variants of persuasiveness, the answer to these questions is captured in a proposition which builds on the following definition and lemma:

**Definition 5.1.6** An argumentation theory  $AT$  is *conflict-free* if no argument in the set of arguments  $S$  of  $AT$  defeats an argument in  $S$ .

Note that this definition is a variant of Definition 3.2.2 of a conflict-free set of arguments which was first developed by Dung in [3].

**Lemma 5.1.7** The set of obviously sceptically (credulously) persuasive elements  $S_{SPE}$  ( $S_{CPE}$ ) of an argument is a subset of the set of attackable elements  $S_{AE}$  of that argument, which makes the number of elements  $|S_{SPE}|$

$(|S_{CPE}|)$  in the former set smaller than or equal to the number of elements  $|S_{AE}|$  in the latter.

**Proof:**

This lemma is easily proven by looking at Definitions 5.1.3 and 5.1.4. We see here that any obviously sceptically (credulously) persuasive element is, in fact, an attackable element.  $\square$

**Proposition 5.1.8** When an argument  $A$  has an obvious sceptical or credulous persuasiveness value of  $v$  for some agent  $p$ , then this means that for  $v * 100\%$  of the attackable elements  $e$  in  $A$  it holds that if  $A$  is defeated by  $p$  on  $e$  in some legal “argue” move  $m$ , then the commitment based argumentation theory which is based on  $p$ ’s commitments after  $m$  is not conflict-free.

**Proof:**

Suppose we have an argument  $A$  which has an obvious sceptical (credulous) persuasiveness value of  $v$  for agent  $p$ . We must now prove that, given this assumption, it holds that for  $v * 100\%$  of the attackable elements  $e$  in  $A$  it holds that if  $A$  is defeated by  $p$  on  $e$  in some legal “argue” move  $m$ , then the commitment based argumentation theory which is based on  $p$ ’s commitments after  $m$  is not conflict-free. In order for us to prove this, we need to look at the definition of persuasiveness again. Our value  $v$  is the result of the ratio  $SPE/AE$  ( $CPE/AE$ ), where  $AE$  is the number of attackable elements of  $A$  and  $SPE$  ( $CPE$ ) the number of obviously sceptically (credulously) persuasive elements of  $A$ . There are two cases now: the case in which  $AE \neq 0$  and the case in which  $AE = 0$ .

- Case 1 ( $AE \neq 0$ ): first, Lemma 5.1.7 tells us that by dividing  $|S_{SPE}|$  ( $|S_{CPE}|$ ) by  $|S_{AE}|$  and multiplying the result by 100, we get the percentage of attackable elements in the argument which are also obviously sceptically (credulously) persuasive elements. This proves how we get from  $v$  to the percentage of attackable elements  $e$  in  $A$  which are also obviously sceptically (credulously) persuasive elements, for which we want to show that if  $A$  is defeated by  $p$  on  $e$  in some legal “argue” move  $m$ , then the commitment based argumentation theory which is based on  $p$ ’s commitments after  $m$  is not conflict-free. Since we treat all types of obviously sceptically (credulously) persuasive elements in the same way in the definitions of persuasiveness, we can just take one of the obviously sceptically (credulously) persuasive elements in  $A$  and

continue the proof with that element. Let us call this obviously sceptically (credulously) persuasive element  $e$ . We now have to prove for this  $e$  that if  $A$  is defeated by  $p$  on  $e$  in some legal “argue” move  $m$ , then the commitment based argumentation theory which is based on  $p$ ’s commitments after  $m$  is not conflict-free. Since  $e$  is an obviously sceptically (credulously) persuasive element for  $p$ , this means that  $p$  can construct, from her commitment based argumentation theory, a justified (defensible) argument  $B$  with conclusion  $e$  if  $e$  is an ordinary premise, or a justified (defensible) argument in which  $e$  is used, if  $e$  is a defeasible inference rule. If, however,  $p$  defeats argument  $A$  on  $e$  in move  $m$ , then this must be done with an argument that has either  $\neg e$  as conclusion if  $e$  is an ordinary premise, or  $\neg n(e)$  as conclusion if  $e$  is a defeasible inference rule. This means that  $p$ ’s own argument  $B$  is also defeated on  $e$  in the commitment based argumentation theory based on  $p$ ’s commitments after  $m$ , which means that this argumentation theory is not conflict-free. This concludes the proof for case 1.

- Case 2 ( $AE = 0$ ): according to Definition 5.1.5,  $v$  is always 1 in this case, which means that 100% of zero attackable elements are obviously sceptically (credulously) persuasive elements. Trivially, since we can say anything about something which does not exist, it follows here that any defeat of  $A$  on a non-existing, attackable element leads to a commitment based argumentation theory which is not conflict-free. This concludes the proof for case 2 and thereby the entire proof.

□

Proposition 5.1.8 is an important result, because it highlights a psychological aspect of the notion of persuasiveness presented in Definition 5.1.5, which can best be described as the feeling one experiences when “one’s own weapon is used against one”. The “weapon” here is an attackable element  $e$  which is sceptically or credulously persuasive to  $p$ , and “using it against” that agent means defeating a move put forward by  $p$  with the help of  $e$ . Of course, an agent may also believe in a number of axioms, which certainly have persuasive power in their own sense, just like attackable elements do. However, this is a different sense than the one captured in Definition 5.1.5, since an argument cannot be defeated on an axiom (for that reason, they are not attackable elements according to Definition 5.1.3), while an argument can be defeated directly on an attackable element which is sceptically or credulously persuasive to some agent  $p$  and contributes to the persuasiveness of an argument by  $p$ ’s dialogue counterpart, but not without this leading to conflicts in the commitments of  $p$ .

In order for us to make a proposition about unobvious sceptical or credulous persuasiveness, we need to look at conflicts between beliefs and commitments. This notion and the proposition about persuasiveness which relies on it are discussed in the next section.

## 5.2 Dishonesty

In [1], three classes of dishonesty are distinguished: *lies*, *bullshit* and *deception*. Since the third of these classes, deception, has to do with intention, which is a concept that we do not address in this thesis, we do not consider it here. Besides lies and bullshit, we will discuss *dialogical* and *logical inconsistency*, which are treated in [7], in this section. Since lies and bullshit are direct violations of two particular members of the set of Grice’s maxims, i.e. respectively “*Do not say what you believe to be false*” and “*Do not say that for which you lack adequate evidence*”, it is interesting for us to investigate them. The other two types of dishonesty that were mentioned, dialogical and logical inconsistency, are interesting because they are not only mentioned in [7], but also in [8] and other literature, as phenomena with which persuasion dialogue systems should be able to deal.

We will use the following assumptions in our examples of the different types of dishonesty: a liberal dialogue system  $DS = (AS, D)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$ , agents  $p$  and  $\bar{p}$ ,  $K_p(p) = \{\neg\varphi, \psi, \neg\psi\}$ ,  $K_p(\bar{p}) = \emptyset$ , a belief based argumentation theory  $BAT(p)$ , empty sets of axioms and defeasible inference rules, a set of strict rules containing all valid inference rules in the classical interpretation of propositional logic and a preference order which gives all elements in its domain the same preference.

### 5.2.1 Lying

In [1], the most simple form of a lie is defined as “some speaker’s utterance of a statement which the speaker knows not to be true”. More complex forms of lies are also possible, such as the type of lie described in [2] which also takes into account a liar’s intent that the hearer will adopt the false belief uttered by the liar. However, since the notion of intention is never used in this thesis, this type of lie is not useful to us.

Let us focus on the simple type of lie. The definition given above contains two terms which do not correspond directly to notions from Prakken’s framework: “knows” and “utterance of a statement”. In the framework, agents have beliefs, some of which, i.e. axioms, have a status resembling that of knowledge, but which are still called “beliefs”. Furthermore, agents in Prakken’s framework utter speech acts which are contained in moves. These

speech acts could be called “statements”, but they do not have truth values of their own. Instead, propositions contained in these speech acts are the entities which possess such truth values. Moreover, a proposition uttered by an agent using a speech act in some move is only relevant to the dialogue if the agent commits to the proposition in that move.

Taking these differences between Caminada’s definition and persuasion dialogues into account, the definition of a lie should become something like: some agent  $p$ ’s move with which  $p$  commits to proposition  $\varphi$  is a lie if  $p$  believes  $\neg\varphi$ . But what does it mean for an agent to believe some proposition? This is where the concepts of a justified or defensible argument and acceptability of an argument’s conclusion come in. We say:  $p$  believes some proposition  $\varphi$  if  $\varphi$  is, either sceptically or credulously, preferred-semantics-acceptable in  $p$ ’s belief based argumentation theory.

We end up with the following definitions of a lie:

**Definition 5.2.1** Given a liberal dialogue system  $DS = (AS, D)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$ , a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$ , a belief based argumentation theory  $BAT(p)$  and attacking move  $m$  by  $p$  which is a legal continuation of  $d$ ,  $m$  is a:

- *hard lie* if  $p$  commits to  $\varphi$  with  $m$  and  $\varphi$  is not credulously preferred-semantics-acceptable in  $BAT(p)$  and  $\neg\varphi$  is sceptically  $S$ -acceptable in  $BAT(p)$ .
- *soft lie* if  $p$  commits to  $\varphi$  with  $m$  and  $\varphi$  is credulously, but not sceptically preferred-semantics-acceptable in  $BAT(p)$  and  $\neg\varphi$  is credulously, but not sceptically  $S$ -acceptable in  $BAT(p)$ .

Algorithms “checkForHardLie” and “checkForSoftLie” could be created, which determine if some move is respectively a hard lie or a soft lie and which both have a complexity of  $C(\text{testAcceptability})$ .

**Example** (hard lie):

Suppose we have the following very simple liberal dialogue:

$p_1: \text{argue } \varphi$

Given the assumptions made at the beginning of this section, move  $p_1$  is a hard lie here, since  $\neg\varphi$  is sceptically preferred-semantics-acceptable in  $BAT(p)$ , whereas  $\varphi$  is not.

**Example** (soft lie):

Suppose we have the following, again very simple, liberal dialogue:

$p_1: \text{argue } \psi$

Given the assumptions made at the beginning of this section, move  $p_1$  is a soft lie here, since both  $\psi$  and  $\neg\psi$  are credulously, but not sceptically preferred-semantics-acceptable in  $BAT(p)$ .

Worthy to note, is the fact that a soft lie cannot be told by an agent with a belief based argumentation theory that has only one preferred extension. We end up with the following proposition:

**Proposition 5.2.2** If the participants of a dialogue  $d$  have belief based argumentation theories that have only one preferred extension, then no soft lies can be displayed in  $d$ .

**Proof:**

Since, in the case of a soft lie, two formulas  $\varphi$  and  $\neg\varphi$  are credulously, but not sceptically acceptable in some agent  $p$ 's belief based argumentation theory  $BAT(p)$ , and we have to prove that this cannot be the case if  $BAT(p)$  has only one preferred extension, the proof for this proposition follows directly from Corollary 3.3.11.  $\square$

Note that hard lies can still be displayed in a dialogue with agents who have belief based argumentation theories that have only one extension. In the case of a hard lie, the proposition that is lied about by some agent is a member of none of the extensions of that agent's belief based argumentation theory, while the negation of that proposition is a member of all extensions. In the case of only one extension, a proposition that is lied about is simply not contained in this extension, while its negation is.

### 5.2.2 Bullshitting

In [1], Caminada defines bullshit as “statements made without the speaker having sufficient knowledge about their validity”. According to Caminada, intentional aspects can also be incorporated in definitions of bullshit. In contrast with the intentional aspects of lying, however, it is, in the case of bullshit, more important that the bullshitter appears knowledgeable to the

hearer about that which is bullshitted about, than that the hearer will adopt it. However, these intentional aspects are not interesting to us, for the same reason as they were not interesting in the case of lies.

Furthermore, a distinction is made in [12] between two types of commitments: those brought about by so-called “concessions” when a dialogue participant surrenders to her opponent and those brought about by “assertives” when a dialogue partner makes a non-surrendering move. This distinction is useful for our understanding of bullshitting, since it is clearly no case of bullshitting when a dialogue participants makes a conceding statement without that participant having sufficient knowledge about its validity, while in the case of an assertive statement a lack of sufficient knowledge does, in fact, make the statement worthy of the label “bullshit”.

Using the same concepts as we used in the definitions of a lie but in a slightly different way, we get the following definition of bullshit.

**Definition 5.2.3** Given a liberal dialogue system  $DS = (AS, D)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$ , a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$ , a belief based argumentation theory  $BAT(p)$  and move  $m$  by  $p$  which is a legal continuation of  $d$ ,  $m$  is called *bullshit* if  $m$  is an attacking move and  $p$  commits to  $\varphi$  with  $m$  and  $\varphi$  is not credulously preferred-semantics-acceptable in  $BAT(p)$  and  $-\varphi$  is not credulously preferred-semantics-acceptable in  $BAT(p)$ .

An algorithm “checkForBullshit” could be created, which determines whether some move is bullshit or not and has a complexity of  $C(testAcceptability)$ .

**Example** (bullshit):

Suppose we have the following liberal dialogue:

$p_1: \text{argue } \chi$

Given the assumptions made at the beginning of this section, move  $p_1$  is bullshit here, since  $p$  is not knowledgeable about  $\chi$ , meaning that neither  $\chi$  nor  $\neg\chi$  is credulously preferred-semantics-acceptable in  $BAT(p)$ .

### 5.2.3 Dialogical incoherence

In [7], Mackenzie introduces the concept of *dialogical incoherence*, which means that a dialogue participant challenges some proposition  $\varphi$  while his commitments imply  $\varphi$ . We will define our own versions of this concept below.

**Definition 5.2.4** Given a liberal dialogue system  $DS = (AS, D)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$ , a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$  and a commitment based argumentation theory  $CAT_d(p)$ ,  $p$  displays:

- *obvious hard dialogical incoherence* with move  $m$  if  $m$  is a legal continuation of  $d$  and  $m$  contains the speech act *why*  $\varphi$  and it holds that  $\varphi$  is sceptically preferred-semantics-acceptable in  $CAT_d(p)$ .
- *obvious soft dialogical incoherence* with move  $m$  if  $m$  is a legal continuation of  $d$  and  $m$  contains the speech act *why*  $\varphi$  and it holds that  $\varphi$  is credulously, but not sceptically preferred-semantics-acceptable in  $CAT_d(p)$ .

Variants are *unobvious* versions of the definitions above, of which the definition can be obtained by replacing the word “obvious” by “unobvious” everywhere in the definition above, and letting  $CAT_d(p)$  be a belief based argumentation theory  $BAT(p)$ .

At first sight, these definitions may seem too strict, since there is a form of challenging, which we call “inquisitive challenging” here, which falls under the definitions but is not improper from an intuitive viewpoint.

Usually, when an agent challenges some proposition, she does this because she wants to defeat the move with which the proposition was put forward. Let us call this way of challenging “aggressive challenging”. When an agent inquisitively challenges some proposition, however, she has other reasons for wanting to find out with what argument her opponent can support the proposition. A reason for an agent to use inquisitive challenging of some proposition is that agent’s doubt about the credibility of her opponent. If the opponent can give an argument for the proposition which is, for example, highly persuasive, then this gives the opponent credibility. However, if the opponent fails to give such a persuasive argument for the proposition, then this does not mean that, in turn, the agent proceeds to attack the opponent. Rather, if no persuasive argument is given as a response to the agent’s inquisitive challenging, the agent may decide that the opponent has low credibility and end the dialogue.

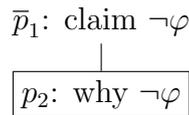
Having elaborated on the concept of inquisitive challenging, we state here that when a agent inquisitively challenges a proposition which follows from her own beliefs or commitments, this is not an attacking move. It is not such a move, because it is unintuitive to say that inquisitive challenges of this kind have defeating power like other types of attacking moves. However, this type of inquisitive challenging does also clearly not belong to the class of surrendering moves. Rather, we state that inquisitively challenging a proposition

which follows from one’s own beliefs or commitments belongs to a separate class of neither attacking nor surrendering, but “neutral” speech acts which do not have defeating power and do not fit into Prakken’s persuasion dialogue framework. Hence, our definitions of dialogical incoherence are not too strict for the framework, but they are, in fact, too strict for a framework in which the discussed third class of “neutral” speech acts is distinguished.

Algorithms “checkForHardDialogicalIncoherence” and “checkForSoftDialogicalIncoherence” could be created, which determine if with some move the player of that move displays respectively hard or soft dialogical incoherence and which both have a complexity of  $C(\text{testAcceptability})$ .

**Example** (unobvious hard dialogical incoherence):

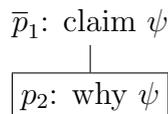
Suppose we have the following liberal dialogue:



Given the assumptions made at the beginning of this section, move  $p_2$  is unobvious hard dialogical incoherence here, since  $p$  challenges  $\neg\varphi$ , while in fact  $\neg\varphi$  is sceptically preferred-semantics-acceptable in  $BAT(p)$ .

**Example** (unobvious soft dialogical incoherence):

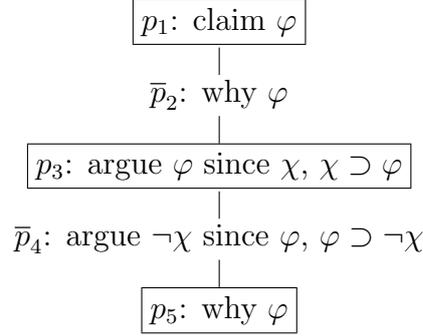
Suppose we have the following liberal dialogue:



Given the assumptions made at the beginning of this section, move  $p_2$  is unobvious soft dialogical incoherence here, since  $p$  challenges  $\psi$ , while in fact both  $\psi$  and  $\neg\psi$  are credulously, but not sceptically preferred-semantics-acceptable in  $BAT(p)$ .

**Example** (obvious hard dialogical incoherence):

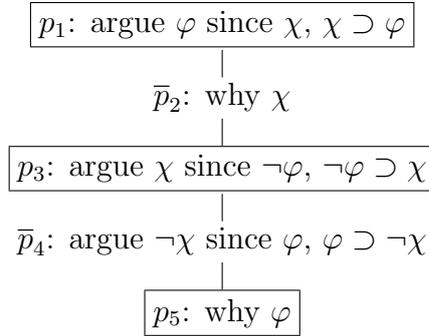
Suppose we have the following liberal dialogue:



Given the assumptions made at the beginning of this section and commitment based argumentation theory  $CAT_{p_1, \bar{p}_2, p_3, \bar{p}_4}(p)$  based on  $C_{p_1, \bar{p}_2, p_3, \bar{p}_4}(p) = \{\varphi, \chi, \chi \supset \varphi\}$ , move  $p_5$  is obvious hard dialogical incoherence here, since  $p$  challenges  $\varphi$  in this move, while in fact  $\varphi$  is sceptically preferred-semantics-acceptable in  $CAT_{p_1, \bar{p}_2, p_3, \bar{p}_4}(p)$ .

**Example** (obvious soft dialogical incoherence):

Suppose we have the following liberal dialogue:



Given the assumptions made at the beginning of this section and commitment based argumentation theory  $CAT_{p_1, \bar{p}_2, p_3, \bar{p}_4}(p)$  based on  $C_{p_1, \bar{p}_2, p_3, \bar{p}_4}(p) = \{\varphi, \chi, \chi \supset \varphi, \neg\varphi, \neg\varphi \supset \chi\}$ , move  $p_5$  is obvious soft dialogical incoherence here, since  $p$  challenges  $\varphi$  in this move, while both  $\varphi$  and  $\neg\varphi$  are credulously, but not sceptically preferred-semantics-acceptable in  $CAT_{p_1, \bar{p}_2, p_3, \bar{p}_4}(p)$ .

Soft dialogical incoherence, both the obvious and the unobvious variant, cannot be displayed by an agent for whom it holds that respectively the commitment or belief based argumentation theory has only one extension. This can be expressed by the following proposition:

**Proposition 5.2.5** If the participants of a dialogue  $d$  have commitment or belief based argumentation theories which have only one preferred extension,

then respectively no obvious or unobvious soft dialogical incoherence can be displayed in  $d$ .

**Proof:**

Since, in the case of obvious or unobvious soft dialogical incoherence, two formulas  $\varphi$  and  $-\varphi$  are credulously, but not sceptically acceptable in respectively some agent  $p$ 's commitment or belief based argumentation theory, and we have to prove that this cannot be the case if this commitment or belief based argumentation theory has only one preferred extension, then the proof for this proposition follows directly from Corollary 3.3.11.  $\square$

Obvious and unobvious hard dialogical incoherence, however, can still be displayed in a dialogue with agents who have belief based argumentation theories that have only one extension. In the case of obvious or unobvious hard dialogical incoherence, some agent challenges a proposition that is a member of all of the extensions of respectively that agent's commitment or belief based argumentation theory. In case there is only one preferred extension, a proposition that is challenged by a move leading respectively to obvious or unobvious hard dialogical incoherence is contained in this extension.

#### 5.2.4 Obvious logical incoherence

In [7], obvious logical incoherence is just called “logical incoherence” and it occurs if there is a proposition  $\varphi$  for which it holds that some dialogue participant's commitment base contains both  $\varphi$  and  $-\varphi$ . In this thesis, an agent displays obvious logical incoherence if she performs a move that makes her commitment based argumentation theory have multiple preferred extensions.

**Definition 5.2.6** Given a liberal dialogue system  $DS = (AS, D)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$  and a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$ ,  $p$  displays *obvious logical incoherence* with move  $m$  in  $d$  if  $m$  is a legal continuation of  $d$  and the commitment based argumentation theory  $CAT_{d,m}(p)$  has multiple preferred extensions.

The algorithm “checkForObviousLogicalIncoherence” could be created, of which the complexity can be expressed in terms of  $C(\text{generateExtensionsFromAT})$  if we design the algorithm such that it first generates all preferred extensions of the input AT and then determines if there are two or more of such extensions. In this case, the complexity of the algorithm is simply equal to  $C(\text{generateExtensionsFromAT})$ .

In the last example of the previous section,  $p$  displays obvious logical incoherence with move  $p_3$ , since she commits to  $\neg\varphi$  in this move while she also committed to  $\varphi$  in  $p_1$ , thereby giving the commitment based argumentation theory  $CAT_{p_1, \bar{p}_2}(p)$  two preferred extensions, i.e.  $\{\varphi\}$  and  $\{\neg\varphi\}$ .

### 5.2.5 Unobvious persuasiveness revisited

We are now ready to present the proposition about unobvious persuasiveness which we promised we would give in Section 5.1.

**Proposition 5.2.7** When an argument  $A$  has an unobvious sceptical or credulous persuasiveness value of  $v$  for some agent  $p$ , then this means that for  $v * 100\%$  of the attackable elements  $e$  in  $A$  it holds that if  $A$  is defeated by  $p$  on  $e$  in some legal “argue” move  $m$  legally continuing some dialogue  $d$ , then if  $e$  is not a defeasible inference rule,  $m$  is a hard or soft lie, and if  $e$  is a defeasible inference rule,  $m$  is hard or soft lie or bullshit, or  $p$ ’s belief based argumentation theory  $BAT(p)$  is not conflict-free.

**Proof:**

Suppose we have an argument  $A$  which has an unobvious sceptical (credulous) persuasiveness value of  $v$  for agent  $p$ . We must now prove that, given this assumption, it holds that for  $v * 100\%$  of the attackable elements  $e$  in  $A$  it holds that if  $A$  is defeated by  $p$  on  $e$  in some legal “argue” move  $m$  which legally continues dialogue  $d$ , then if  $e$  is not a defeasible inference rule,  $m$  is a hard or soft lie, and if  $e$  is a defeasible inference rule,  $m$  is a hard or soft lie or bullshit, or  $p$ ’s belief based argumentation theory  $BAT(p)$  is not conflict-free. Given that  $AE$  is the number of attackable elements of  $A$ , we have two cases:  $AE \neq 0$  and  $AE = 0$ .

- Case 1 ( $AE \neq 0$ ): Lemma 5.1.7 tells us that by dividing  $|S_{SPE}| (|S_{CPE}|)$  by  $|S_{AE}|$  and multiplying the result by 100, we get the percentage of attackable elements in the argument which are also unobviously sceptically (credulously) persuasive elements. This proves the relationship between  $v$  and the percentage of attackable elements  $e$  in  $A$  for which we want to give our proof. We take an unobviously sceptically (credulously) persuasive element  $e$  from  $A$  and continue the proof with that element. We now have to prove for this  $e$  that if  $A$  is defeated by  $p$  on  $e$  in some legal “argue” move  $m$ , then if  $e$  is not a defeasible inference rule,  $m$  is a hard or soft lie, and if  $e$  is a defeasible inference rule,  $m$  is a hard or soft lie or bullshit, or  $p$ ’s belief based argumentation theory

$BAT(p)$  is not conflict-free. Since  $e$  is an unobviously sceptically (credulously) persuasive element for  $p$ , this means that  $p$  can construct, from her belief based argumentation theory, a justified (defensible) argument with as conclusion  $e$ , if  $e$  is not a defeasible inference rule, or a justified (defensible) argument in which  $e$  is used, if  $e$  is a defeasible inference rule. If, however,  $p$  defeats argument  $A$  on  $e$  in move  $m$  with an argument  $C$ , then  $Conc(C) = -e$ , if  $e$  is not a defeasible inference rule, or  $Conc(C) = -n(e)$ , if  $e$  is a defeasible inference rule. This means that  $p$  commits to  $-e$  or  $-n(e)$ , which, in turn, implies that  $m$  is a hard or soft lie if  $e$  is not a defeasible inference rule.  $m$  is a hard lie in the case where  $v$  is an unobvious sceptical persuasiveness value and  $-e$  is not credulously acceptable in  $p$ 's belief based argumentation theory under some semantics  $S$ , while  $e$  is sceptically  $S$ -acceptable in this belief based argumentation theory.  $m$  is a soft lie in the case where  $v$  is an unobvious credulous persuasiveness value, and both  $e$  and its negation are credulously, but not sceptically  $S$ -acceptable in  $p$ 's belief based argumentation theory. In the case where  $e$  is a defeasible inference rule,  $p$  can construct from her belief base a justified (defensible) argument  $D$  that used  $e$ , while at the same time she commits to  $-n(e)$  in the attacking move  $m$ . There are three possible scenario's here. In the first scenario,  $n(e)$  is sceptically (credulously)  $S$ -acceptable in  $BAT(p)$ , which means that  $m$  is a hard (soft) lie. In the second scenario,  $-n(e)$  is sceptically (credulously)  $S$ -acceptable in  $BAT(p)$ , which means that  $BAT(p)$  is not conflict-free, since the argument  $D$  is attacked by an argument with  $-n(e)$  as conclusion. In the third scenario, neither  $n(e)$  nor its negation are credulously  $S$ -acceptable in  $BAT(p)$ , which means that  $m$  is bullshit. This concludes the proof for case 1.

- Case 2 ( $AE = 0$ ): according to Definition 5.1.5,  $v$  is always 1 in this case, which means that 100% of zero attackable elements are unobviously sceptically (credulously) persuasive elements. Trivially, since we can say anything about something which does not exist, it follows here that any move defeating  $A$  on a non-existing, attackable element is a hard or soft lie. This concludes the proof for case 2 and thereby the entire proof.

□

The psychological aspect of the notion of persuasiveness, which was discussed in Section 5.1 and highlighted by Proposition 5.1.8, is also highlighted by Proposition 5.2.7, since, in the case of an argument which is unobviously persuasive to an agent, that agent either displays improper behaviour or

turns out to have a conflict in her beliefs if she defeats that argument on one of its unobviously persuasive elements.

### 5.3 Irrelevance

Before we can define irrelevant moves, we first have to ask ourselves: when is a move irrelevant? Since it is a dialogue agent’s goal to win a persuasion dialogue, it seems natural to say that a move by some agent is irrelevant if that move does not contribute to the agent’s goal of winning the dialogue. But now, the question becomes: when does a move contribute to an agent’s goal of winning a dialogue and when does it not?

In [8], Prakken defines two types of dialogue systems which assure, in their own ways, that only relevant moves are legal in the system and irrelevant moves are not. He calls these types of dialogue systems “strongly relevant” and “weakly relevant” dialogue systems. In strongly relevant dialogue systems, a move is legal only if it changes the status of the original claim. In weakly relevant dialogue systems, a move by some agent is legal only if it either adds a so-called ‘winning part’ for that agent to the dialogue, or removes a winning part of the opponent from the dialogue.

**Definition 5.3.1** Let  $d$  be a liberal dialogue with participants  $p$  and  $\bar{p}$  currently won by player  $p$ . A winning part  $d_p$  of  $d$  is recursively defined as follows.

- First include  $m_1$ ;
- for each move  $m$  of  $p$  that is included, if  $m$  is surrendered, include all its surrendering replies, otherwise include all its attacking replies;
- for each attacking move  $m$  of  $p$  that is included, include one attacking reply  $m'$  that is *in* in  $d$ .

The algorithm “countWinningParts” could be created, of which the complexity  $C(\text{countWinningParts})$  is  $O(n \times m)$ , where  $n$  is the number of leaf moves in the input dialogue tree  $d$  made by agent  $p$  for which the algorithm counts the winning parts, and  $m$  is the number of move nodes in  $d$ . This is the complexity, since each leaf move of  $p$  can be part of mostly one winning part, which is a fact that follows trivially from the definition of a winning part, and for each of the  $n$  leaf moves the algorithm has to run, in the worst case scenario, through all  $m$  nodes of the entire dialogue tree in order for it to be able to determine whether the move belongs to a winning part or not.

We can now take Prakken’s definitions of relevance and use them to define irrelevant moves.

**Definition 5.3.2** Given a liberal dialogue  $d$  with participants  $p$  and  $\bar{p}$ , an attacking move  $m$  is a *strongly irrelevant move* if  $m$  is a legal continuation of  $d$  and  $m$  does not change the status of the original claim in  $d$ . A surrendering move is strongly irrelevant iff its attacking counterparts are strongly irrelevant.

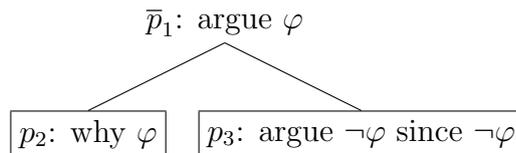
The algorithm “checkForStrongIrrelevance” could be created, of which the complexity is  $O(1)$ , since no matter what the size of any of the data structures fed as input to the algorithm is, checking if the status of the original claim has changed always takes the same amount of time.

**Definition 5.3.3** Given a liberal dialogue  $d$  with participants  $p$  and  $\bar{p}$ , an attacking move  $m$  by participant  $p$  is a *weakly irrelevant move* if  $m$  is a legal continuation of  $d$  and with  $m$ ,  $p$  neither adds a winning part for herself to  $d$  nor removes a winning part for her opponent from  $d$ . A surrendering move is weakly irrelevant iff its attacking counterparts are weakly irrelevant.

The algorithm “checkForWeakIrrelevance” could be created, of which the complexity is equal to  $C(\text{countWinningParts})$ , since the algorithm needs to count the number of winning parts for the input agent twice, which is a constant number of times, and compare the results, which takes a constant amount of time.

**Example** (strongly irrelevant move):

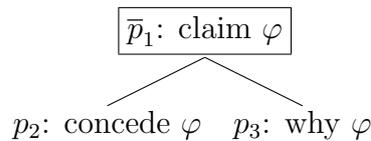
Suppose we have the following liberal dialogue:



Move  $p_3$  is a strongly but not weakly irrelevant move here, since does not change the status of the initial move, but does create a new winning part for  $p$ .

**Example** (weakly irrelevant move):

Suppose we have the following liberal dialogue:



Move  $p_3$  is a weakly irrelevant move here, since it does not create a new winning part for  $p$  nor takes away a winning part of  $\bar{p}$ .

## 5.4 Verbosity

In Section 3.3, definitions for two types of arguments are given: general arguments, which are just called ‘arguments’ in the definition, and minimal arguments. If an argument is not minimal, then this means that it is the opposite, i.e. non-minimal. A non-minimal argument contains premises which are not necessary for the inference expressed by the argument. Hence, these premises are superfluous, meaning that verbosity is displayed by an agent using a non-minimal argument.

**Definition 5.4.1** Given a liberal dialogue  $d$  with participants  $p$  and  $\bar{p}$ ,  $p$  uses a non-minimal argument in  $m$  if  $m$  is a legal continuation of  $d$  and  $m$  contains the speech act ‘argue  $A$ ’, where  $A$  is not a minimal argument.

The algorithm “checkForNonMinimalArgument” could be created, of which the complexity is  $O(n \times (2^m - 1))$ , where  $n$  is the number of subarguments of the input argument and  $m$  is the number of premises of the strict inference rule used in the argument which has the greatest number of premises of all strict inference rules used in the argument. This makes the worst case scenario that in which all strict inference rules have the same number of premises. This is the complexity of the algorithm, because in order to determine whether an argument is minimal or not, the algorithm needs to visit all subarguments of the argument once, and for each subargument of the form  $A_1, \dots, A_m \rightarrow \varphi$  it has to check for at most  $2^m - 1$  strict subsets  $\{a_1, \dots, a_i\}$  of the set  $Conc(\{A_1, \dots, A_m\})$  if the strict rule  $a_1, \dots, a_i \rightarrow \varphi$  exists.

## 5.5 Improper behaviour unrelated to individual moves

So far, we have only looked at types of improper behaviour of agents which are related to particular moves. It is, however, also possible that improperness resides in the belief base of an agent. This type of improperness is necessarily present at both the beginning and the end of a dialogue, since, under the assumption that belief bases are static, it cannot come into existence during a dialogue as the result of an agent’s move.

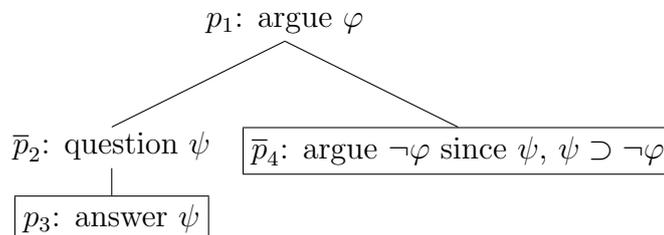
**Definition 5.5.1** Given a dialogue system  $DS = (AS, D)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$ , a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$  and a belief based argumentation theory  $BAT(p)$ ,  $p$  displays *unobvious logical incoherence* in  $d$  if  $BAT(p)$  has multiple preferred extensions.

Obviously, the complexity of an algorithm “checkForUnobviousLogicalIncoherence” would be the same as that for *checkForObviousLogicalIncoherence* which was given in Section 5.2.4.

## 6 Variation 1: inquiry

An inquiry speech act for persuasion dialogues is introduced by Mackenzie in [7], and also plays a part in Fulda’s [4], which is about cross-examination of witnesses before a trial. The speech act could be translated into the following question in natural language: “do you believe, disbelieve or have no belief at all about  $\varphi$ ?”. The speech act *question*  $\varphi$  has three possible attacking replies: *answer*  $\varphi$ , meaning ‘I believe  $\varphi$ ’; *answer*  $\neg\varphi$ , meaning ‘I believe  $\neg\varphi$ ’; and *ignorant*  $\varphi$ , meaning ‘I believe nor disbelieve  $\varphi$ ’.

Let us illustrate the use of this speech act with an example. Suppose we have two agents,  $p$  and  $\bar{p}$ .  $p$  wants to persuade  $\bar{p}$  of  $\varphi$ , so she starts a liberal dialogue with  $\bar{p}$  by arguing for  $\varphi$  in the first move.  $\bar{p}$  could attack  $p$ ’s initial argument now with the argument  $A = \neg\varphi$  since  $\psi, \psi \supset \neg\varphi$ . Since  $p$  is committed to neither one of the formulas in  $A$ , she could, after  $\bar{p}$  has placed  $A$ , attack the premises in  $A$  again without introducing a conflict in her own commitments. But suppose now that  $P$  were, in fact, committed to both  $\psi$  and  $\psi \supset \neg\varphi$ . In that case,  $\bar{p}$ ’s argument would be more persuasive, since none of its premises could be attack by  $p$  without her introducing a conflict in her own commitments. Even if  $p$  were just committed to one of the premises in  $A$ , this would make  $A$  stronger. Thus, it is of tactical advantage for  $\bar{p}$  to make  $p$  commit to at least one of the premises in  $A$ , before she actually uses the argument. This is where the *question* speech act comes in. Suppose  $\bar{p}$  asks *question*  $\psi$  here and  $p$  answers that she believes  $\psi$ . Now,  $\bar{p}$  can argue for  $\neg\varphi$  without having to worry that  $p$  will attack the premise  $\psi$  in  $A$ , such as is illustrated as follows:



## 6.1 Extension of the communication language

Let us extend the communication language table defined in Section 3.5 with the speech acts introduced in this section. This yields the following new communication language:

Acts	Attacks	Surrenders
claim $\varphi$	why $\varphi$ question $\psi$	concede $\varphi$
why $\varphi$	argue A ( $\text{Conc}(A) = \varphi$ )	retract $\varphi$
argue A	why $\varphi$ ( $\varphi \in \text{Prem}(A)$ ) argue B (B defeats A)  question $\varphi$	concede $\varphi$ ( $\varphi \in \text{Prem}(A)$ or $\varphi = \text{Conc}(A)$ )
question $\varphi$	answer $\varphi$ answer $-\varphi$ ignorant $\varphi$	
answer $\varphi$	why $\varphi$	concede $\varphi$
ignorant $\varphi$		
concede $\varphi$		
retract $\varphi$		

The following commitment rules are drawn up for the new speech acts (below  $s$  denotes the speaker of the move).

- If  $s(m) = \text{question}(\varphi)$  then  $C_s(d, m) = C_s(d)$ ;
- If  $s(m) = \text{answer}(\varphi)$  then  $C_s(d, m) = C_s(d) \cup \{\varphi\}$ ;
- If  $s(m) = \text{ignorant}(\varphi)$  then  $C_s(d, m) = C_s(d)$ .

## 6.2 Improper behaviour related to inquiry

### 6.2.1 Pretending to be ignorant

This form of dishonesty is the opposite of bullshitting: instead of telling her opponent something about which she is not knowledgeable, an agent keeps some piece of information about which she is, in fact, knowledgeable to herself when she pretends to be ignorant.

**Definition 6.2.1** Given a liberal dialogue system  $DS = (AS, D)$  with the communication language specified in Section 6.1 and the new commitment rules of Section 6.1 added to those of liberal dialogues, argumentation theory

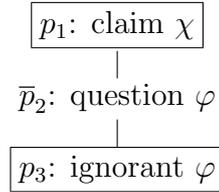
$AT = (AS, \mathcal{K}, \preceq)$ , a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$ , a belief based argumentation theory  $BAT(p)$  and move  $m$  by  $p$  which is a legal continuation of  $d$  and contains the speech act ‘*ignorant  $\varphi$* ’,  $p$  is:

- *pretending to be ignorant in a hard manner* if  $\varphi$  is sceptically preferred-semantics-acceptable in  $BAT(p)$  or  $\neg\varphi$  is sceptically  $S$ -acceptable in  $BAT(p)$ .
- *pretending to be ignorant in a soft manner* if  $\varphi$  is credulously, but not sceptically preferred-semantics-acceptable in  $BAT(p)$  or  $\neg\varphi$  is credulously  $S$ -acceptable in  $BAT(p)$ .

Algorithms for checking if an agent pretends to be ignorant in a hard or soft manner could be created, which have a complexity of  $C(\text{testAcceptability})$ .

**Example** (pretending to be ignorant in a hard manner):

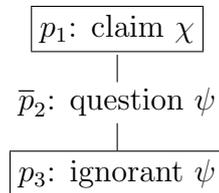
Suppose we have the following liberal dialogue:



Given the assumptions made at the beginning of Section 5.2,  $p$  is pretending to be ignorant in a hard manner by making move  $p_3$ , since  $p$  is not ignorant about  $\varphi$  because  $\neg\varphi$  is sceptically preferred-semantics-acceptable in  $BAT(p)$ .

**Example** (pretending to be ignorant in a soft manner):

Suppose we have the following, again very simple, liberal dialogue:



Given the assumptions made at the beginning of Section 5.2,  $p$  is pretending to be ignorant in a soft manner by making move  $p_3$ , since both  $\psi$  and  $\neg\psi$  are credulously, but not sceptically preferred-semantics-acceptable in  $BAT(p)$ .

### 6.2.2 Asking an irrelevant question

A question asked by  $p$  is considered irrelevant here if the question cannot serve to make some argument, which  $p$  can construct from her beliefs and commitments, more persuasive to her opponent.

**Definition 6.2.2** Given a liberal dialogue system  $DS = (AS, D)$  with the communication language specified in Section 6.1 and the new commitment rules of Section 6.1 added to those of liberal dialogues, argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$ , a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$ , a commitment based argumentation theory  $CAT_d(p)$ , semantics  $S$  for interpreting Dung-style argumentation frameworks and a question  $q$  by  $p$  which is a legal continuation of  $d$ ,  $q$  is *sceptically relevant* iff there is an argue speech act  $a = \text{'argue } A\text{'}$  for which it holds that:

- there is a move containing  $a$  which is a legal continuation of  $d$  and has an obvious sceptical persuasiveness for  $\bar{p}$  of  $SP$ ,
- the continuation of  $d$  by  $q$  leads to the new dialogue  $d'$ ,
- there is a reply to  $q$  which, by continuing dialogue  $d'$ , leads to dialogue  $d''$ , and
- there is a move containing  $a$  which is a legal continuation of  $d''$  and has an obvious sceptical persuasiveness for  $\bar{p}$  of  $SP' > SP$ .

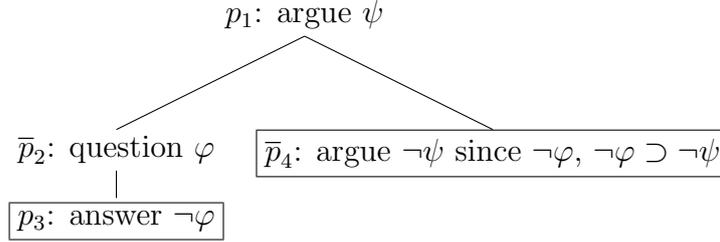
The definition of a *credulously relevant* question can be obtained by taking the definition above and replacing the word “sceptically” with “credulously” everywhere in the definition, “sceptical” with “credulous”, “ $SP$ ” with “ $CP$ ” and “ $SP'$ ” with “ $CP'$ ”.

One can imagine an algorithm for determining whether a question is relevant in some way or not, which first generates all arguments that can be constructed from the input commitment or belief based argumentation theory by converting that argumentation theory to an argumentation framework using *convertATtoAF*. After this, the algorithm determines, for every generated argument, if one of the three possible replies to the input question makes the argument more persuasive, which means that the algorithm also has to calculate the persuasiveness of some kind of all generated arguments. Thus, the complexity of the algorithm is  $C(\text{convertATtoAF}) + n \times (2 \times C(\text{calcPersuasiveness}))$ .

**Definition 6.2.3** A question is *sceptically irrelevant* if it is not sceptically relevant. Likewise, a question is *credulously irrelevant* if it is not credulously relevant.

**Example** (question which is neither sceptically nor credulously irrelevant):

Suppose we have the following liberal dialogue:



Given the assumptions made at the beginning of Section 5.2,  $\bar{p}$ 's move  $\bar{p}_4$  is neither a sceptically nor a credulously irrelevant question, since there is, in fact, an argue speech act  $a = \text{'argue } A\text{'}$ , with  $A = \neg\chi$  since  $\neg\varphi, \neg\varphi \supset \neg\chi$ , for which it holds that:

- $\bar{p}_2$  could have been a different move containing 'argue  $A$ '. In that case,  $\bar{p}_2$  would have had an obvious sceptical persuasiveness of 0 for  $p$ , since  $p$ 's commitment base after the initial claim contains only  $\chi$  and therefore none of the three attackable elements in  $A$  are sceptically preferred-semantics-acceptable in the commitment based argumentation theory  $CAT_{p_1}(p)$ . Also,  $\bar{p}_2$  would have had an obvious credulous persuasiveness of  $\frac{2}{3}$  for  $p$ , since two of the three attackable elements in  $A$ , i.e.  $\neg\chi$  and  $\neg\varphi \supset \neg\chi$ , are credulously preferred-semantics-acceptable in  $CAT_{p_1}(p)$ ; and
- the move actually containing 'argue  $A$ ' in the dialogue, move  $\bar{p}_4$ , has an obvious sceptical persuasiveness of  $\frac{1}{3}$  for  $p$ , since  $p$ 's commitment base after  $p_3$  contains  $\chi$  and  $\neg\varphi$  which means that one of the three attackable elements in  $A$ , i.e.  $\neg\varphi$ , is sceptically preferred-semantics-acceptable in the commitment based argumentation theory  $CAT_{p_1, \bar{p}_2, p_3}(p)$ . This value is higher than in the case where  $\bar{p}_2$  would have contained 'argue  $A$ ', meaning that  $\bar{p}$ 's question in move  $\bar{p}_2$  has increased the obvious sceptical persuasiveness of  $A$ . Also,  $\bar{p}_4$  has an obvious credulous persuasiveness of 1 for  $p$ , since all attackable elements in  $A$  are credulously preferred-semantics-acceptable in  $CAT_{p_1, \bar{p}_2, p_3}(p)$ . This value is also higher than in the case where  $\bar{p}_2$  would have contained 'argue  $A$ ',

meaning that  $\bar{p}$ 's question in move  $\bar{p}_2$  has also increased the obvious credulous persuasiveness of  $A$ .

### 6.2.3 Asking the same question more than once

Since we assumed that belief bases are static, it is considered an irrelevant move when an agent asks the same question more than once in a dialogue.

**Definition 6.2.4** Given a liberal dialogue  $d$  with participants  $p$  and  $\bar{p}$ , based on a liberal dialogue system with the communication language specified in Section 6.1 and the new commitment rules of Section 6.1 added to those of liberal dialogues,  $p$  asks the same question more than once by placing question  $q$  in  $d$  if  $q$  is a legal continuation of  $d$  and  $q$  contains the speech act ‘question  $\varphi$ ’, which is also contained in another move already made by  $p$  in  $d$ .

In the worst case scenario, an algorithm for testing the uniqueness of a question in a dialogue would have to visit every node, minus the initial node, since that node cannot be a question, and the node holding the input question, in the input dialogue tree in order for it to determine whether some question is asked twice or not, which makes the complexity of the algorithm  $O(n - 2)$ , where  $n$  is the number of nodes in the input tree.

## 7 Variation 2: indications of incoherence

### 7.1 Exploiting logical incoherence: *resolve* $\varphi, -\varphi$

This speech act is introduced by Mackenzie in [7] and is generalized here. When a player displays obvious logical incoherence, the other player may demand that the first retracts one of her commitments with the aim of resolving that incoherence. This speech act can then be attacked again by ‘resolved  $\varphi$ ’ or ‘resolved  $-\varphi$ ’ after the attacker has resolved the logical incoherence in her part of the dialogue, leaving respectively  $\varphi$  or  $-\varphi$  sceptically preferred-semantics-acceptable in her commitment based argumentation theory and making respectively  $-\varphi$  or  $\varphi$  not even credulously preferred-semantics-acceptable in that argumentation theory. Such an attack using the ‘resolved’ speech act is necessary after an agent has restored logical coherence in the dialogue, since if she does not use this attack, the move by her opponent containing the ‘resolve’ speech act remains undefeated, even though the move has no actual attacking power any more because it indicates an inconsistency which does not exist any more.

Let us, again, extend the communication language table defined in Section 3.5 with the speech acts introduced in this subsection. This yields the following new communication language:

Acts	Attacks	Surrenders
claim $\varphi$	why $\varphi$	concede $\varphi$
why $\varphi$	argue A ( $\text{Conc}(A) = \varphi$ )	retract $\varphi$
argue A	why $\varphi$ ( $\varphi \in \text{Prem}(A)$ ) argue B (B defeats A) resolve $\varphi, -\varphi$ ( $\varphi \in \text{Prem}(A)$ )	concede $\varphi$ ( $\varphi \in \text{Prem}(A)$ ) or $\varphi = \text{Conc}(A)$ )
resolve $\varphi, -\varphi$	resolved $\varphi$ resolved $-\varphi$	
resolved $\varphi$		
concede $\varphi$		
retract $\varphi$		

The following commitment rules are drawn up for the new speech acts (below  $s$  denotes the speaker of the move).

- If  $s(m) = \text{resolve}(\varphi, -\varphi)$  then  $C_s(d, m) = C_s(d)$ ;
- If  $s(m) = \text{resolved}(\varphi)$  then  $C_s(d, m) = C_s(d) - \{\varphi\}$ .

Now, we can identify the following types of improper behaviour related to logical incoherence:

- If an agent  $p$  in some dialogue  $d$  uses in the move legally continuing  $d$  the speech act *resolve*  $\varphi, -\varphi$  while it is not the case that both  $\varphi$  and  $-\varphi$  are credulously acceptable in the commitment based argumentation theory of  $\bar{p}$  under some semantics, then  $p$  is dishonest and thus she displays improper behaviour.
- If an agent  $p$  in some dialogue  $d$  uses in the move legally continuing  $d$  the speech act *resolved*  $\varphi$  while  $\varphi$  and  $-\varphi$  are still both credulously acceptable in the commitment based argumentation theory of  $p$  under some semantics, then  $p$  is dishonest and thus she displays improper behaviour.

## 7.2 Exploiting dialogical incoherence: *eo ipso* $\varphi$

When one party challenges  $\varphi$  while her commitments actually imply  $\varphi$  (dialogical incoherence), the ‘*eo ipso*’ speech act may be used by the other party

to demand that the challenger either retracts one of her implying commitments or concedes  $\varphi$ . This speech act can then be attacked again by ‘*non eo ipso*  $\varphi$ ’, after the attacker has solved the dialogical incoherence in her part of the dialogue. Just like in the case of the ‘resolved’ speech act, an attack using the ‘*non eo ipso*’ speech act is necessary after an agent has restored dialogical coherence in the dialogue, since if she does not use this attack, the move by her opponent containing the ‘*eo ipso*’ speech act remains undefeated, even though the move has no actual attacking power any more because it indicates an inconsistency which does not exist any more.

Let us, also here, extend the communication language table defined in Section 3.5 with the speech acts introduced in this subsection. This yields the following new communication language:

Acts	Attacks	Surrenders
claim $\varphi$	why $\varphi$	concede $\varphi$
why $\varphi$	argue A ( $\text{Conc}(A) = \varphi$ ) eo ipso $\varphi$	retract $\varphi$
argue A	why $\varphi$ ( $\varphi \in \text{Prem}(A)$ ) argue B (B defeats A)	concede $\varphi$ ( $\varphi \in \text{Prem}(A)$ or $\varphi = \text{Conc}(A)$ )
eo ipso $\varphi$	non eo ipso $\varphi$	
non eo ipso $\varphi$		
concede $\varphi$		
retract $\varphi$		

The following commitment rules are drawn up for the new speech acts (below  $s$  denotes the speaker of the move).

- If  $s(m) = \text{eo\_ipso}(\varphi)$  then  $C_s(d, m) = C_s(d)$ ;
- If  $s(m) = \text{non\_eo\_ipso}(\varphi)$  then  $C_s(d, m) = C_s(d) - \{\varphi\}$ .

Now, we can identify the following types of improper behaviour related to dialogical incoherence:

- If an agent  $p$  in some dialogue  $d$  uses in the move legally continuing  $d$  the speech act *eo ipso*  $\varphi$  while  $\varphi$  is not credulously or sceptically acceptable in the commitment based argumentation theory of  $\bar{p}$  under some semantics, then  $p$  is dishonest and thus she displays improper behaviour.

- If an agent  $p$  in some dialogue  $d$  uses in the move legally continuing  $d$  the speech act *non eo ipso*  $\varphi$  while  $\varphi$  is still credulously or sceptically acceptable in the commitment based argumentation theory of  $p$  under some semantics, then  $p$  is dishonest and thus she displays improper behaviour.

## 8 Definitions of improper behaviour

We are now ready to precisely define improper behaviour for a specific class of dialogue systems called “liberal\* dialogue systems”, which are liberal dialogue systems in which the communication language defined in Section 6.1 (variation 1) is adopted and the commitment rules specified in Section 6.1 are added to the commitment rules of liberal dialogue systems. The speech acts defined in Section 7 are not adopted, since they are used to indicate inconsistencies which we actually wish to ban altogether from our dialogues. Therefore, those speech acts are useless to us.

### 8.1 Definitions without non-optimal usefulness

Since we have given no definitions for non-optimally useful behaviour yet, we first present definitions of improper behaviour which build on the definitions of types of improper behaviour we did see already.

**Definition 8.1.1** If  $m \in P(d)$ , then  $m$  *obviously violates* the maxim of

- *honesty* if  $m$  is:
  - a move with which the speaker of  $m$  displays obvious hard or soft dialogical incoherence, or
  - a move with which the speaker of  $m$  displays obvious logical incoherence;
- *relevance* if  $m$  is:
  - a weakly irrelevant move, or
  - a sceptically or credulously irrelevant question, or
  - a move with which the speaker of  $m$  asks the same question more than once in  $d$ ;
- *briefness* if  $m$  is a move containing a non-minimal argument.

An algorithm “checkForObviousMaximViolation” can be conceived that needs to check all of the conditions of Definition 8.1.1 in sequence, meaning that its complexity is obtained by adding up the complexities of checking the conditions of Definition 8.1.1.

**Definition 8.1.2** If  $m \in P(d)$ , then  $m$  *unobviously violates* the maxim of

- *honesty* if  $m$  is:
  - a hard lie, or
  - a soft lie, or
  - bullshit, or
  - a move with which the speaker of  $m$  displays unobvious hard or soft dialogical incoherence, or
  - a move with which the speaker of  $m$  pretends to be ignorant in a hard or soft manner.

One can imagine an algorithm “checkForUnobviousMaximViolation” which needs to check all of the conditions of Definition 8.1.2 in sequence, meaning that its complexity is obtained by adding up the complexities of checking the conditions Definition 8.1.2.

## 8.2 Defining different types of non-optimal usefulness

### 8.2.1 Premature surrendering

An agent may be presented with many opportunities to surrender during a dialogue. Sometimes surrendering is an optimally useful move, but in other cases an agent surrenders to some move while in fact she could have attacked it instead. We call this type of surrendering “premature surrendering” and declare it non-optimally useful since an agent who does her best to win a dialogue will not display this kind of behaviour in that dialogue. Note that in some cases surrendering is optimally useful, even when an attacking move can be made instead. These are cases in which an agent attacks her opponent with an argument, which, in turn, the opponent can defeat again with an argument which she can construct from her commitments and which has an obvious sceptical persuasiveness of 1. In these cases, an attack leads to nothing.

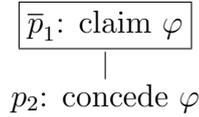
**Definition 8.2.1** Given a dialogue system  $DS = (AS, D)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$ , a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$ , participant  $p$  displays *premature surrendering* with surrendering move  $m$  if

1.  $m$  is a legal continuation of  $d$  which is not weakly irrelevant, and
2.  $p$  can construct an attacking move  $m'$  on the basis of  $AT$  which is also a legal continuation of  $d$ , and
3.  $m'$  does not make  $p$  obviously or unobviously violate any maxim for persuasion dialogues, and
4.  $\bar{p}$  cannot construct an argument with an obvious sceptical persuasiveness of 1 from her commitment based argumentation theory  $CAT_{d,m}(\bar{p})$  that defeats  $m'$ .

An algorithm “checkForPrematureSurrendering” can be conceived that needs to test all four conditions of Definition 8.2.1 in sequence, which makes the complexity equal to  $C(\text{convert}AT\text{to}AF) + C(\text{checkForWeakRelevance}) + C(n) + C(\text{checkForObviousMaximViolation}) + C(\text{checkForUnobviousMaximViolation}) + n \times C(\text{calcPersuasiveness})$ , where  $n$  is the number of arguments generated by  $\text{convert}AT\text{to}AF$  when it is called by the algorithm.

**Example** (premature surrendering):

Suppose we have the following liberal dialogue with participants  $p$ , which has belief base  $K_p(p) = \{\neg\varphi\}$ , and  $\bar{p}$ , which has belief base  $K_p(\bar{p}) = \{\varphi\}$ :



Agent  $p$  displays premature surrendering with move  $p_2$  here, since the first condition of Definition 8.2.1 is met because this move is not weakly irrelevant (its attacking counterparts, which are rebuttals or undercutters of  $\bar{p}_1$  on  $\varphi$ , take away a winning part of  $\bar{p}$ ). Also the second and third conditions are met since  $p$  could have attacked  $\bar{p}_1$  in  $p_2$  using an argument  $A$  with  $\neg\varphi$  as conclusion, which would not have obviously or unobviously violated any maxim. In addition, the final condition of Definition 8.2.1 is met since  $\bar{p}$  would not have been able to construct an argument with an obvious sceptical persuasiveness of 1 from her commitment based argumentation theory  $CAT_{\bar{p}_1,p_2}(\bar{p})$  that defeated  $A$ .

### 8.2.2 Attacking with a non-most persuasive argument

Another type of behaviour which does not contribute optimally to an agent’s goal of winning a dialogue, is if an agent attacks her opponent with an

argument in some move, while she could also have attacked the opponent with a different, more persuasive argument in that same move.

**Definition 8.2.2** Given a dialogue system  $DS = (AS, D)$ , argumentation theory  $AT = (AS, \mathcal{K}, \preceq)$ , a dialogue  $d$  of  $DS$  with participants  $p$  and  $\bar{p}$ , participant  $p$  attacks  $\bar{p}$  with a *sceptically non-most persuasive argument* in move  $m$  if  $m$  is a legal continuation of  $d$  containing an 'argue' speech act and  $m$  has an obvious sceptical persuasiveness for  $\bar{p}$  of  $SP$ , and  $p$  can construct an argument  $m'$  on the basis of  $AT$  with a obvious sceptical persuasiveness for  $\bar{p}$  of  $SP' > SP$ , which is also a legal continuation of  $d$  and does not make  $p$  obviously or unobviously violate any maxim for persuasion dialogues.

The definition of *attacking with a credulously non-most persuasive argument* can be obtained by taking the definition above and replacing the word “sceptically” with “credulously” everywhere in the definition, “sceptical” with “credulous”, “ $SP$ ” with “ $CP$ ” and “ $SP'$ ” with “ $CP'$ ”.

An algorithm “checkForNonMostPersuasiveArgument” can be conceived which has a complexity of  $C(\text{calcPersuasiveness}) + C(\text{convertATtoAF}) + n \times (C(\text{calcPersuasiveness}) + C(\text{checkForObviousMaximViolation}) + C(\text{checkForUnobviousMaximViolation}))$ , where  $n$  is the number of arguments generated by  $\text{convertATtoAF}$  when it is called by the algorithm.

### 8.3 Definition including non-optimal usefulness

Finally, we can define improper behaviour in its totality.

**Definition 8.3.1** [Improper behaviour] If  $m \in P(d)$ , then  $m$  is improper behaviour if  $m$  obviously or unobviously violates any of the maxims for persuasion dialogues contained in Definitions 8.1.1 and 8.1.2, or if, by using  $m$ , the speaker of  $m$  displays premature surrendering or attacks her opponent with a non-most persuasive argument.

## 9 Banning improper behaviour from persuasion dialogues

In this section, two ways of banning improper behaviour from liberal\* dialogue systems are discussed here: one using protocol rules and one using guidelines for agent design. These two ways will collectively be denoted using the term “instruments for ensuring proper behaviour”. The reason that we need two ways of banning of improper behaviour, is that we would like

to respect public semantics in our dialogues by doing so. In Section 4.1, we assumed that an agent’s belief base and set of defeasible inference rules are private, which means that under public semantics, no outside observer, including any protocol rule, can know their contents. Other objects, such as commitment bases, are assumed to be public, meaning that their contents are visible to outside observers such as protocol rules under public semantics. Since the definitions of particular types of improper behaviour refer to the contents of private objects, these types of improper behaviour cannot be banned from dialogues using protocol rules. Therefore, we need another way to deal with these types of improper behaviour. One way of doing so, is creating guidelines that specify how agents can be designed in such a way that they do not display those types of improper behaviour which cannot be banned using protocol rules.

## 9.1 Guidelines for agent design

The following guidelines are given to the designer of an agent  $p$  which will be participating in liberal\* dialogue systems:

*Design  $p$  such that  $p$  never displays unobvious logical incoherence in a dialogue and that a move by  $p$  in some liberal\* dialogue  $d$  never unobviously violates any maxim, or is a move with which  $p$  displays premature surrendering or attacks her opponent with a sceptically or credulously non-most persuasive argument.*

These guidelines, however, can be further refined, since they contain redundancy. This redundancy is caused by the fact that agent designers are advised to design agents such that they do not display unobvious logical incoherence. Since this means that agents designed according to these guidelines have belief based argumentation theories for which it holds that these theories only have one extension under preferred semantics, Corollary 3.3.11 tells us that the notions of sceptical and credulously acceptability of conclusions in argumentation theories become interchangeable. This, in turn, means that the following types of improper behaviour, which rely on formulas being credulously, but not sceptically acceptable in an argumentation theory under preferred semantics, cannot occur any more:

- soft lies,
- soft dialogical inconsistency, and
- pretending to be ignorant in a soft manner.

Taking into account also that there is no distinction any more between the different types of attacking with a non-most persuasive argument, we can improve the guidelines presented above by cutting these types of improper behaviour out of the definition, which leads to the following new guidelines:

*Design  $p$  such that  $p$  never displays unobvious logical incoherence in a dialogue and that a move by  $p$  in some liberal\* dialogue  $d$  is never:*

- *a move with which  $p$  displays premature surrendering, or*
- *a move with which  $p$  attacks her opponent with a sceptically non-most persuasive argument, or*
- *a hard lie, or*
- *bullshit, or*
- *a move with which the speaker of  $m$  displays unobvious hard dialogical incoherence, or*
- *a move with which the speaker of  $m$  pretends to be ignorant in a hard manner.*

## 9.2 Protocol rules

Under the assumption that the guidelines for agent design given in Section 9.1 are followed, a new rule is added to the protocol for liberal\* dialogue systems. Here, redundancy is kept out of the protocol rule by not taking into account types of improper behaviour involving credulous acceptability of some formula in an argumentation theory under particular semantics, which have a sceptical variant.

If  $m \in P(d)$ , then:

- $R_8$ :  $m$  is not:
  - a move with which the speaker of  $m$  displays obvious hard dialogical incoherence, or
  - a move with which the speaker of  $m$  displays obvious logical incoherence, or
  - a weakly irrelevant move, or
  - a sceptically irrelevant question, or

- a move with which the speaker of  $m$  asks the same question more than once in  $d$ , or
- a move containing a non-minimal argument.

An algorithm “checkRuleR8” for checking if protocol rule  $R_8$  holds at some point during a dialogue is conceivable of which the complexity can be obtained by first adding up the complexities of obtaining all the data structures needed for checking the different cases of the rule, and then adding to the result of that addition the complexity of each individual check.

### 9.3 Justification of the instruments for ensuring proper behaviour

In the previous subsections, two instruments for ensuring proper behaviour were presented: guidelines for agent designers and protocol rule  $R_8$ . The question that arises now, is: can the instruments for ensuring proper behaviour be justified? We will provide justification for the instruments for ensuring proper behaviour by showing that the instruments possess the following properties:

**Effectiveness** There is no case in which the instruments for ensuring proper behaviour cannot be applied.

**Desirability of consequence** If an agent is forced to surrender in a dialogue by the instruments for ensuring proper behaviour, then this is desirable.

The effectiveness of the instruments for ensuring proper behaviour can be proven by showing that no *hard dialogical paradox* can occur in liberal\* dialogue systems.

**Definition 9.3.1** Given a liberal\* dialogue  $d$  and the set of all maxims for persuasion dialogues  $C$ , a *hard dialogical paradox* occurs if for every move  $m$  which is a legal continuation of  $d$ , it holds that  $m$  violates one of the maxims  $c \in C$ .

**Proposition 9.3.2** No hard dialogical paradox can occur in any liberal\* dialogue  $d$ .

**Proof:**

In order to prove Proposition 9.3.2, we need to show that, given any dialogue  $d = m_1, \dots, m_i, \dots, m_n$ , there is always a move legally continuing  $d$  which does not violate a maxim. We prove this by showing for each of the different types of speech acts  $\mathbf{a}$  in our communication language, that if  $s(m_i)$  is of type  $\mathbf{a}$ , then there is a move  $m_{n+1}$  with  $t(m_{n+1}) = m_i$  legally continuing  $d$  which does not violate a maxim.

$\mathbf{a} \in \{claim(\varphi), why(\varphi), argue(A), answer(\varphi)\}$  : in all of these cases, it holds that if  $m_{n+1}$  is a surrendering move (either of the ‘retract’ kind, in the case of *why*  $\varphi$ , or of the ‘concede’ kind, in the other cases), then  $m_{n+1}$  is either a move with which the speaker of that move displays premature surrendering, or not. In the first case, Definition 8.2.1 says that there is an attacking move legally continuing  $d$  which does not obviously or unobviously violate any maxim. In the second case,  $m_{n+1}$  is either a move which does not violate a maxim, or weakly irrelevant, since weak irrelevance is, besides premature surrendering, the only type of improper behaviour that can be caused by a surrendering move. In order for a weakly irrelevant surrendering move to be possible, its target must already have been surrendered to in an earlier move  $m_j$ , which means that  $s(m_{n+1}) = s(m_j)$ , which makes  $m_{n+1}$ , in fact, an illegal continuation of  $d$  since protocol rule  $R_4$ , which was introduced in Section 3.4, says that if two moves in a dialogue have the same target, then the speech acts they contain must be different. Here, we end up at a contradiction, and therefore  $m_{n+1}$  is a surrendering move legally continuing  $d$  which does not violate a maxim.

$\mathbf{a} = question(\varphi)$  : since unobvious logical incoherence cannot occur in liberal\* dialogue systems, it is the case that for any formula  $\varphi$ , an agent’s belief base either contains  $\varphi$  or  $-\varphi$ , or contains neither one of these formulas. Thus, to any “question  $\varphi$ ” move, there is always exactly one legal reply that the speaker of  $m_{n+1}$  can give which does not violate a maxim. This finishes our proof.

□

This proof shows that the instruments for ensuring proper behaviour are effective because one type of dialogical paradox, i.e. the hard type, cannot occur in liberal\* dialogues. There is, however, another type of dialogical paradox, which we call a “soft dialogical paradox”. This type can, in fact, occur in liberal\* dialogues.

**Definition 9.3.3** Given a liberal\* dialogue  $d$  and the set of all maxims for persuasion dialogues  $C$ , a *soft dialogical paradox* occurs if for every attacking

move  $m$  which is a legal continuation of  $d$ , it holds that  $m$  violates one of the maxims  $c \in C$ .

A simple example which shows us that soft dialogical paradoxes can occur in liberal\* dialogues, is a dialogue in which one of the participants attacks the other with an argument  $A$  that targets an ordinary premise and has an unobvious sceptical persuasiveness of 1, and there are no other moves in the dialogue which can be attacked. In this case,  $A$  cannot be attacked or challenged, since this would respectively lead to hard lying or unobvious hard dialogical incoherence. Furthermore, no question can be legally asked in this case, since there is no argument which can be made more persuasive by means of such a question. Thus, a soft dialogical paradox occurs in the dialogue and the participant attacked by  $A$  is forced to make a surrendering move.

The fact that a dialogue agent can be forced by the instruments for ensuring proper behaviour to surrender in a dialogue, makes it meaningful for us to demonstrate that this consequence of the instruments is desirable. This can be done by answering the following question: can it be justified that there are cases in which an agent is forced by the instruments for ensuring proper behaviour to surrender in a dialogue?

We can answer this question by connecting the application of the instruments for ensuring proper behaviour to the concept of improper behaviour. We distinguish between proper and improper dialogue agent behaviour in the first place because the first type of behaviour is desirable, while the second is not. Because improper behaviour is undesirable, it is desirable that the instruments for ensuring proper behaviour keep agents from displaying it. Since this precisely what they do, we can conclude that it is desirable that dialogue agents can be forced by the instruments for ensuring proper behaviour to surrender in a dialogue.

By showing that the instruments for ensuring proper behaviour possess both the property of effectiveness and the property of desirability of consequence, we have justified the instruments for ensuring proper behaviour.

## 10 Conclusion

In this thesis, a number of behavioural conventions for agents participating in argumentation based persuasion dialogues have been introduced, which were inspired on Grice's conversational maxims. Using these conventions as guidelines, different types of improper behaviour of persuasion dialogue participants have been defined and discussed, and example algorithms for

determining whether some type of improper behaviour has occurred at some point during a dialogue, and their complexities have been presented for these types. Also, two different methods of banning improper behaviour from persuasion dialogues have been developed and justified.

We started this thesis with three research questions, which are answered as follows:

- [*How participants of persuasion dialogues should behave ideally*] In Section 4 we answered this question by using Grice’s conversational maxims as inspiration for our own set of maxims for argumentation based persuasion dialogues. These maxims define exactly those conventions for which it holds that if a dialogue agent adheres to all of them in a dialogue, she behaves ideally in that dialogue.
- [*What types of improper behaviour of dialogue participants can occur in Prakken’s framework, depending on the design choices made*] In Sections 5, 6 and 7 we answered this question by formally defining a number of types of improper behaviour in dialogue systems with different communication languages.
- [*How these types of improper behaviour can be banned from dialogues which are modelled by Prakken’s framework*] In Section 9 we answered this question by defining a specific type of dialogue system called a liberal\* dialogue system, for which we devised a protocol rule which bans certain types of improper behaviour from this type of system. As for the types of improper behaviour which we could not ban from liberal\* dialogue systems using this protocol rule, we devised guidelines for agent designers which enables them to design their agents such that they do not display these types of improper behaviour in liberal\* dialogue systems.

The relevance of the results of the research presented in this thesis to the scientific field of AI is threefold. First, the thesis bridges the gaps between literature on the subjects of argumentation based persuasion dialogue systems (notably Prakken’s work [8]), dishonesty (notably Caminada’s work [1]) and the language-philosophical discipline of pragmatics (Grice’s influential work [5]) by combining concepts from these different pieces of literature in a new way and building further on them. Second, definitions presented in this thesis such as those for persuasiveness enrich the conceptual jargon of formal argumentation theory which is used in works such as [3] and [10]. Third, the results of this thesis are not only theoretically but also practically interesting, since they bring us closer to the realization of actual software agents

participating in persuasion dialogues which behave in a desirable manner.

As an extension of this research, it could be investigated what types of improper behaviour occur in instantiations of Prakken's framework for which (configurations of) design choices are made that have not been discussed in this thesis. For example: other speech acts could be introduced or a single-move turntaking rule could be adopted. Also, one could investigate what types of improper behaviour occur in other persuasion dialogue framework, or in dialogue systems which do not model persuasion. The outcomes of such research could be compared to the outcomes of this thesis. Finally, an interesting topic of investigation is how the guidelines presented in Section 9.1 and the protocol rules presented in Section 9.2 can be most efficiently implemented in respectively dialogue agents and dialogue systems.

## References

- [1] Martin Caminada. Truth, lies and bullshit; distinguishing classes of dishonesty. In *Proceedings SS@IJCAI 2009 (Workshop on Social Simulation)*, pages 39–50, 2009.
- [2] Roderick M Chisholm and Thomas D Feehan. The intent to deceive. *The Journal of Philosophy*, 74(3):143–159, 1977.
- [3] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.
- [4] Joseph S Fulda. The logic of “improper cross”. *Artificial Intelligence and Law*, 8(4):337–341, 2000.
- [5] H Paul Grice. Logic and conversation. *Syntax and Semantics*, 3:41–58, 1975.
- [6] Diana Grooters. Paraconsistent logics in argumentation systems. Master’s thesis, Universiteit Utrecht, 2014.
- [7] Jim D Mackenzie. Question-begging in non-cumulative systems. *Journal of Philosophical Logic*, 8(1):117–133, 1979.
- [8] Henry Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15(6):1009–1040, 2005.
- [9] Henry Prakken. Formal systems for persuasion dialogue. *The Knowledge Engineering Review*, 21(02):163–188, 2006.
- [10] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.
- [11] Douglas N Walton. *Logical Dialogue-games*. University Press of America, Lanham, Maryland, 1984.
- [12] Douglas N Walton and Erik CW Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Press, Albany, New York, 1995.