What Should We Do?:

Computational Representation of Persuasive Argument in Practical Reasoning

Thesis submitted in accordance with the requirements of the
University of Liverpool for the degree of
Doctor in Philosophy
by
Katie Marie Atkinson

September 2005

To my parents, with love and gratitude.

Abstract

“What Should We Do?” by Katie Atkinson

The design and development of autonomous software agents requires a multitude of elements to be considered and accounted for. In order for a software agent to be considered ‘intelligent’ it must be able to perform effective reasoning about its beliefs and the environment in which it is situated, and also act in this environment. It must therefore reason both about what is the case, and what should be done: the latter is known as practical reasoning. Additionally, it must also be able to interact and reason with other such agents in its environment, as it may rely on them for information and help to enable it to accomplish its tasks. This thesis is concerned with one particular aspect of such agency: modelling the process of argument in practical reasoning to equip autonomous agents with the capability to determine the best action to take, in a given situation. The background setting for this work deals with the topic of practical reasoning and attempts to address some issues regarding its treatment in philosophy, as well as the problems inherent in the computational modelling of such reasoning. The main output of the study is a theory of persuasion in practical reasoning which makes use of techniques from the field of argumentation theory, to enable autonomous software agents to construct and reason about arguments in support of and against proposals for action. The theory is embodied in a model describing how agents based on the Belief-Desire-Intention (BDI) architecture can put forward a proposal for action and how this proposal can be systematically attacked in a variety of ways. This enables them to consider all available options and come to a conclusion about the best action to take, in the given context. The underlying theory extends a well established account from the field of philosophy, based on the use of argument schemes and critical questions. The account given is then formalised in terms to enable its representation in agent systems.

The underlying theory has formed the basis for a number of applications: an implementation of a dialogue game protocol to provide a proof of concept; an implementation to provide computer mediated support for human decision making in a particular context; and finally, a formalism to enable autonomous agents to reason about decisions regarding actions. The account for use in BDI agents is applied to three example domains – law, medicine and politics – to show how BDI agents can reason and argue about matters of practical action, in accordance with the theory.

Acknowledgements

Firstly, I would like to thank all who have provided me with financial assistance. My PhD research has been funded by the Engineering and Physical Sciences Research Council (EPSRC) and I am most grateful for this support. Additionally, the Department of Computer Science at the University of Liverpool, the Faculty of Science at the University of Liverpool and AgentLink III have also provided me with financial assistance to enable me to attend numerous conferences and present my work. I am grateful to all parties for this assistance.

Secondly, I would like to thank my academic colleagues. I am grateful to all members of the Department of Computer Science at the University of Liverpool who have taken an interest in my work and helped me throughout the course of my PhD by attending my talks, providing me with feedback and engaging in discussions with me about my research. In particular, I would like to thank Frans Coenen, Paul Dunne, Floriana Grasso, Ullrich Hustadt, Wiebe van de Hoek, Dave Jackson, Steve Phelps and Michele Zito. I first became interested in AI and multi-agent systems after having studied these topics on modules for my bachelor's degree. In particular, I attended courses delivered by Mike Wooldridge and Simon Parsons (who is now at the Department of Computer and Information Science, Brooklyn College, USA) and they gave stimulating, knowledgeable and enthusiastic lectures on these subjects. I would like to thank them both for initially sparking my interest in the subject and I feel that it is an honour to have been taught by them. I would also like to thank my fellow PhD students at Liverpool with whom I have shared the trials and tribulations of our PhD programs (as well as a few good parties). In particular, thanks go to Alison Chorley with whom I shared an office, as well the peaks and troughs of our studies. Outside of Liverpool there are numerous people who have helped me by sending me papers, answering my questions, reviewing my work, listening to my talks and commenting on my work. I want to thank all who have shown an interest in my work and development. Thanks go to some people in particular who have provided me with help and engaged in discussions with me about my work: Rogier van Eijk of the Department of Information and Computing Sciences, University of Utrecht, Tom Gordon of Fraunhofer FOKUS, Berlin, Sanjay Modgil of the Advanced Computation Laboratory at Cancer Research UK, Andrea Omicini of

DEIS, Alma Mater Studiorum, University of Bologna, Henry Prakken of the Department of Information and Computing Sciences, University of Utrecht, Iyad Rahwan of the School of Informatics, British University of Dubai, Chris Reed of the Department of Applied Computing, University of Dundee, and Simon Wells of the Department of Applied Computing, University of Dundee. Thanks also go to my thesis examiners, Paul Leng of the Department of Computer Science at the University of Liverpool and Douglas Walton of the Department of Philosophy at the University of Winnipeg, Canada, for interesting and enjoyable discussions about my work during my viva voce examination.

Thirdly, the support I have received from my family and friends throughout has been overwhelming. My parents, Carmel and Ken Greenwood, have shared all the joyful and testing times with me and they have always been there to help me through any difficult patches in any way that they could. I cannot express how very grateful I am to them both for the love and support they have given to me and for the opportunities that I have been presented with which stem from this. Great thanks are also due to my ‘other family’, Linda and Richard Atkinson, for their constant interest, advice and encouragement during my PhD research. My extended family and friends have also been extremely supportive and shown an interest in my academic development. I am particularly grateful to two people for their friendship and support: Louisa Clarke and Jasmine Kitses, who have both been invaluable friends to me. There is one more family member who has been steadfastly by my side throughout the whole of my PhD studies: my husband Sam Atkinson. In addition to sharing the good times with me, his patience in listening to my worries and his help in supporting me through the hard times have been never-ending. Sam, words cannot express how much your love, support and sacrifice have meant to me throughout this time and always. And just for the record, I will finally be getting a job now my PhD is finished!

Finally, there are two people to whom I am more grateful than I can convey: my supervisors Peter McBurney and Trevor Bench-Capon. I could not have hoped to have been supervised by two finer people, on both an academic and a personal level. I have learnt so much from them both through engaging in many stimulating discussions with them and they have always provided me with invaluable feedback about my work. I have benefitted greatly from having their combined knowledge at my disposal whenever I had any queries and their doors were always open for me to come and discuss any issues I that had. This support, and the guidance they have provided, has helped me immeasurably. Peter, Trevor, it has been a pleasure and an honour to have been supervised by you both and I will miss our meetings where new ideas were debated, countless stories were retold and many laughs were had. You have both contributed so much towards making my time as a PhD student such a valuable and rewarding experience, and for this I am forever indebted to you both.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"The first duty of a wise advocate is to convince
his opponents that he understands their arguments
and, sympathises with their just feeling."
Samuel Taylor Coleridge, English Poet, (1772–1834).

In this chapter I outline my research question, provide a general overview of the topic that my research falls under, and give an outline of the structure of this thesis.

1.1 Research Question

There are many research challenges to be met if we are to realise societies of autonomous software entities. In order for software programs to become truly autonomous we must equip them with the capabilities to reason about their beliefs regarding the world they inhabit, and the actions that they can take in order to further their design objectives, the latter being known as "practical reason". This is, of course, an enormous task and the work presented in this thesis focuses on one particular part of the latter requirement. The fundamental problem addressed in this thesis is:

By what means may autonomous software agents make, question, defend and jointly reason about proposals for action?

This question arose from consideration of how we are able to persuade and convince others to change their minds about commitment to actions through argumentative techniques and how we can model this in software agents. In my search for answers to this question I draw upon a number of disciplines including: computer science, artificial intelligence, mathematics, philosophy, linguistics, law and politics, as will be seen in the forthcoming chapters.

1.2 Overview of Topic

One of the aims of artificial intelligence (AI) is to enable the construction of useful distributed computer systems that are equipped with the capabilities to display intelligence and make decisions autonomously. One branch of AI that is attempting to deal with these issues is ‘multi-agent systems’. Over the last two decades this field has received considerable attention from researchers who are addressing a wide variety of the issues of AI in an attempt to build systems of distributed, autonomous software agents. Although the field is now firmly established within computer science there is still no universally accepted definition of an agent. The definition that I am following in this thesis is taken from Wooldridge [169], which is itself adapted from an earlier definition by Wooldridge and Jennings [170]:

> “An *agent* is a computer system that is *situated* in some *environment*, and that is capable of *autonomous action* in this environment in order to meet its design objectives.” [169, p. 15]

In addition to this, a collection of such agents situated together in an environment and capable of interacting with one another is known as a ‘multi-agent system’. In order for these entities to be deemed as intelligent, there are a number of capabilities that we would expect such agents to possess, again taken from a list defined by Wooldridge and Jennings [170] and described in [169]:

> **Reactivity**: Intelligent agents are able to perceive their environment, and respond in a timely fashion to changes that occur in it in order to satisfy their design objectives.
>
> **Proactiveness**: Intelligent agents are able to exhibit goal-directed behaviour by *taking the initiative* in order to satisfy their design objectives.
>
> **Social ability**: Intelligent agents are capable of interacting with other agents (and possibly humans) in order to satisfy their design objectives.

Equipping agents with these capabilities is no easy task and before we can address the issue of how to construct mechanisms for reasoning and interaction amongst agents, we first need to provide solid theories upon which the design of agents can be based. Such theories will need to supply us with, amongst other things, mechanisms that will enable agents to reason with information about their environment, beliefs and actions.

The successful creation and deployment of intelligent agents has a wealth of implications for numerous application areas. I describe a few of these application areas below.

The Internet has had an unprecedent impact upon the way business is conducted and web access has also become ubiquitous amongst personal users in the western world. This has introduced the concept of electronic commerce (eCommerce) [87] which has revolutionised the ways in which people shop and conduct business. Although the Internet is still in its relative infancy it is a vast territory in which AI and agent technologies can be exploited. New Internet technologies are currently being developed in which multi-agent systems are being used for both commercial and business purposes. However, these new technologies also bring with them a whole host of new social and technological issues that need to be considered. These include legal issues as agents begin to enter into contracts and perform transactions on their own authority.

The emergence of web technologies has led to the computerisation of numerous ‘traditional’ business processes in the public, as well as the private, sector. The Government’s ability to communicate with the public through online provisions has led to the emergence of a new method of governance: eDemocracy. This area has the potential to dramatically alter the way in which governments rule and interact with the public. Not only has the Internet widened visibility and access to information for members of the public, but it has also enabled a whole range of dynamic processes to be implemented electronically. The exchange of information between citizens and ministers has altered from the traditional methods of letter writing and discussions at council meetings to email contact and web interactions. Not only do such processes have the potential to save governments time and money in the long run, but they also encourage participation from citizens in this new ‘online age’. The transformation of democracy into an electronic medium is currently making great advances, though again this field is still in its infancy and there are many more advances to be made in future research in this area. Numerous countries are engaged in the trial and development of new interactive systems for eDemocracy, such as those for e-voting [152]. However, one important issue that must be addressed in such systems, and in eDemocracy provisions in general, is that of trust and security. Nonetheless, proposals for new systems for eGovernment are attempting to meet this aim, e.g., [5, 132]. Thus, with the introduction of safe and efficient web-based services governments have the opportunity to exploit the benefits of new computer technologies, including AI techniques and agent systems, to provide accessible, efficient and useful systems through which democracy can be effectively conducted.

Another important field that has benefitted greatly from the integration of new computer technology is medicine. With regard to AI, one of the most notable experiments for integrating AI within medicine was the MYCIN project conducted at Stanford University [41]. The output of this project was an expert system to aid doctors in the diagnosis and recommendation of treatment for patients with bacterial infections of the blood. MYCIN was constructed using a number of AI techniques, though it is not classified as an autonomous agent due to its lack of proactive and social abilities.

However, in recent years there have been a growing number of projects researching the integration of autonomous agent systems into medical applications. One particular example of a group who are working on such projects is the Advanced Computation Lab at Cancer Research UK who are investigating how multi-agent systems can be used for simulation, decision making, communication and risk assessment purposes, among others [62, 63, 86].

The examples I have given above are just a few of the many potential fields that AI and multi-agents systems technologies can provide benefit to, and they range from applications aimed at individuals, to large-scale complex applications. There are numerous additional areas in which the application of multi-agent systems is currently being investigated. This suggests that there is great potential for the field of multi-agent systems to contribute to solving some of the problems presented by AI and provide effective and useful new technology.

In addition to the numerous application areas of multi-agent systems there are also a number of different methods and models upon which autonomous agents can be constructed. The model that the account in this thesis is tailored for use in is the popular Belief-Desire-Intention model, as will be discussed in Section 2.3. However, before going on to examine my research topic within the context of multi-agent systems, I will first examine the background theories that are needed to enable us to develop such agents.

To shed some light upon how we might go about producing mechanisms to build autonomous agents, we can turn to the field of philosophy to give us insights into how intelligent behaviour is manifest in humans. As stated in the definition of my research question, the type of reasoning that this thesis is concerned with is that of action – what to do in a given situation – so here I look to the field of practical reasoning in philosophy to begin my investigations. Numerous philosophers dating from recent times and going back to the time of the ancient Greek Philosophers, have given in depth analyses of the nature of practical reasoning. Practical reasoning embodies a number of distinctive and interesting features that have been described and accounted for in a variety of different ways. To date there is no universally agreed standard for the representation and treatment of practical reasoning. Thus, there are however, a number of accounts to turn to for insights into how we can best treat practical reasoning to enable its effective representation in software agents.

Part of the process of practical reasoning involves considering all options available to us in any given situation. There may be competing options that require careful consideration to enable the best decision to be taken. A sub field of philosophy that can help in dealing with this issue is argumentation theory. This field is concerned with the presentation, interaction and evaluation of arguments that support or reject a particular position on a matter. Argumentation is an important part of critical reasoning as it enables the evaluation of different perspectives and points of view to be considered, which facilitates the critical evaluation of the issues in question. Argumentation provides an arena in which consistency and logic can be reasoned about. It is also extremely useful in situations where knowledge is incomplete or inconsistent. It will often be the case that autonomous agents must operate in environments where knowledge is incomplete and where agents have a number of actions available to execute, from which they must choose one. Hence, a section of the multi-agent systems community has turned its attention to this field to try to deal with some of the issues associated with the effective design of agent reasoning mechanisms. The work presented in this thesis turns to argumentation theory in an attempt to model argumentative techniques that can be deployed by agents in scenarios where the competing reasons for justifying an action must be critically considered and evaluated. The method from this field that I will focus on is the concept of argument schemes and critical questions, which provide us with a precise structure by which we can present and criticise justifications for action.

In addition to having coherent reasoning mechanisms, agents also need to be able to communicate effectively with their counterparts, in order to fulfil their social ability requirements. Again, we can turn to philosophy, this time to the field of linguistics, to gain understanding of how this is achieved in humans and how this may in turn help us to build effective methods for agent communication. Human communication consists of a multitude of complex and varied rules of interaction. By examining the basic building blocks upon which languages are comprised we can gain insight into the necessary elements that are needed to build languages for communication between intelligent agents. Construction of a new language that is expressive enough to be useful, whilst at the same time being concise enough to be used by artificial entities, is no easy task. Proposals for solutions to this problem have been given, but a standardised language for use in intelligent agents is a huge goal for the multi-agent systems community to achieve, though significant research is being dedicated to the development of this area, as will be discussed in Section 2.6.

To summarise, the field of multi-agent systems offers exciting prospects for the development of technology that can be of benefit to a wide variety of disciplines. There are indeed numerous problems and challenges that need to be addressed in the quest for the successful creation of artificially intelligent entities. The field draws upon a number of interdisciplinary topics to address all the different aspects needed for the construction, integration and application of intelligent agents. In particular, philosophy has much to give to the theoretical foundations upon which multi-agent systems are built. The account that follows in this thesis presents a theory of persuasion in practical reasoning based upon a philosophical representation using argument schemes and critical questions. This theory is then taken forward and transformed into a dialogue game protocol for human mediation. Finally the account is formalised for representation in BDI agents to enable such agents to reason about justifications for action in a persuasion setting. This approach is applied to three different application domains.

Given the above areas that I have identified as the setting for my research contribution, I hereby outline the main goals of this thesis:

1. To provide a theory of persuasion within the setting of practical reasoning which accounts for the defeasible nature of reasoning about action.
2. Following Searle's account of rationality in action, to separate the objective and subjective components within my theory to provide explanation of how and why rational disagreement can and does occur in practical reasoning.
3. To show how my theory of persuasion over action can be transformed into a computational account that can be effectively deployed in autonomous software agent systems. By using the theory software agents should be able to recognise the objective and subjective components involved in the reasoning, and respond to criticisms appropriately.
4. To provide theoretical examples of how my computational account can be used by BDI agents in a number of different domains which involve reasoning about actions. The examples should show that it is important to model subjectivity in practical arguments, and in doing this, my computational account should effectively model human practical reasoning, more so than traditional decision theoretic accounts.

In the conclusions I present in Chapter 10 I will re-visit these research goals to discuss how well they have been met. I now outline the structure of this thesis.

1.3 Thesis Structure

This thesis is structured into ten chapters as follows:

Chapter 1 is this chapter in which I have defined my research question.

Chapter 2 presents a literature survey of existing research which is relevant to the contributions I present in this thesis.

Chapter 3 introduces a theory of persuasion over action which extends an existing account of practical reasoning from philosophy.

Chapter 4 proposes a protocol for dialogues of persuasion over action and it takes the theory articulated in Chapter 3 as its underlying model. This chapter also gives a brief discussion of an implementation of the protocol for use in human mediated dialogues.

Chapter 5 takes the theory of persuasion over action forward for use in BDI agents. This is done through the provision of definitions to articulate how a BDI agent can put forward and attack a justification for action. This chapter then explains how the agent can choose the best action to commit to, from the justifiable set of actions, through use of an abstract method of argumentation.

Chapter 6 presents two example applications which use the theory of persuasion over action in an eDemocracy setting. The first application is an implemented online mediation system to allow the public to express their views on a government issue. The second application addresses the same issue but uses the definitions provided in Chapter 5 to show how real-life debates about political actions can be undertaken by BDI agents based upon my account.

Chapter 7 presents a second example application showing how BDI agents can reason about decisions for action in the medical domain, in accordance with the account presented in Chapter 5.

Chapter 8 presents a third and final example application from the legal domain to show how a well known legal case can be represented and reasoned about by BDI agents, according to the account presented in Chapter 5.

Chapter 9 provides a summary of the three example applications and discusses the individual and interesting features of each.

Chapter 10 is the final chapter and it provides a summary of all the work presented in the thesis, as well as a discussion of possible avenues for future research.

I also include three additional elements as appendices. **Appendix A** presents an outline of a denotational semantics for the dialogue game protocol, named PARMA, detailed in Chapter 4 and this semantics was developed in joint work with Peter McBurney. The denotational semantics is intended to supplement the axiomatic semantics of PARMA presented in Chapter 4 and as it is only an outline in need of further development, I include it as an appendix. **Appendix B** contains further details of the specification, design, testing and evaluation of the Java program which embodies the PARMA Protocol. The material in Appendix B is intended to supplement the descriptions and discussions of the implementation of the protocol given in Chapter 4. **Appendix C** contains the definitions for Dung's Argumentation Frameworks [55] and Bench-Capon's Value-Based Argumentation Frameworks [26]. Both these frameworks are abstract methods for evaluating the status of arguments and they are discussed in Section 2.5 and in further detail in Section 5.5. Additionally, the method presented in this thesis to enable BDI agents to reason about actions in accordance with my theory also makes use of Value-Based Argumentation Frameworks. Thus the definitions provided in Appendix C are included for reference purposes.

Some of the work presented in this thesis has been developed jointly with other co-authors to whom I am most grateful for their help and contributions. I have also presented parts of this work at various refereed conferences, workshops and seminars and I thank all reviewers and audiences for their comments and suggestions. Segments of work presented in this thesis have been published, or accepted for publication, as joint work with my supervisors Trevor Bench-Capon and Peter McBurney as follows:

- The tables in Section 2.4.4, which model how shifts can occur between dialogues from the Walton and Krabbe typology [167], have been published in [74][1].

- The theory of persuasion over action presented in Chapter 3 appears in [12], [17], [19] and [75].

- Chapter 4 presents a protocol named PARMA which extends the theory of persuasion over action from Chapter 3 and the details of the protocol and an implementation of it are published in [10], [11], [15] and [19].

- The representation of the theory of persuasion over action for use in BDI agents as described Chapter 5 has been published in [9], [14] and [16].

- Chapter 6 presents two examples of the application of the theory of persuasion to the political domain. The first example describes an implementation of a mediation system for eGovernment that has been published in [13]. The second example is a theoretical application in the same domain which makes use of the BDI agent application detailed in Chapter 5 and this example has been published in [17] and [18].

- A second example application using the medical domain for a setting is given in Chapter 7 and this is based on joint work with Sanjay Modgil, of the Advanced Computation Laboratory at Cancer Research UK, and Trevor bench-Capon and this appears in [20].

- A third and final example application using the setting of the legal domain is presented in Chapter 8 and this is based on work that has been published in [14] and [27].

[1] The work in [74] and [75] was published under my previous surname of ‘Greenwood’.

Chapter 2

Literature Review

In this chapter I present an overview of the existing research literature that is relevant to the issues addressed in this thesis. Section 2.1 begins with a discussion of practical reasoning in philosophy. Section 2.2 examines the work of John Searle who makes a number of interesting observations about the nature of practical reasoning. These observations are foundational to the theory of practical reasoning I present in Chapter 3. Section 2.3 turns towards practical reasoning in autonomous agents. Here I give a brief reprisal of the use of, and problems associated with, practical reasoning in the field of autonomous agents and multi-agent systems and in particular, in the Belief-Desire-Intention agent architecture. Section 2.4 surveys the area of argumentation theory in philosophy and discusses a well known typology of dialogue classifications. Section 2.5 extends the discussions of the previous section to show how argumentation has recently been applied to the discipline of artificial intelligence, and in particular multi-agent systems. Section 2.6 gives a summary of the use of dialogue games in philosophy and computer science, and a discussion of the communication methods used in agent systems. Finally, Section 2.7 concludes with a summary of the chapter.

2.1 Practical Reasoning

In this section I examine the topic of practical reasoning, from its early philosophical roots to its treatment in more recent literature. I look at a number of definitions and examples of this form of reasoning and I discuss some of the features and problems inherent in it.

2.1.1 Overview

Practical reasoning is reasoning about what should be done according to some criterion: what is morally correct, what is most pleasurable, what is most prudent on financial or health grounds, and so on. These and other criteria often compete with each other and what *should* be done is always relative to a particular agent, in a particular situation, from a particular perspective. Despite the fact that such reasoning occurs on a common basis in the conduct of activities in the everyday life of most people, this type of reasoning has not been studied within computer science or philosophy nearly as extensively as has reasoning about beliefs.

Much research in AI has focused on mechanisms to enable artificial entities to reason about beliefs about the world. AI traditionally however, involves more than this. Since its earliest days it has also been concerned with artefacts capable of acting so as to modify their environment. Indeed, it could be argued that intelligence requires such an ability: that intelligence can only be manifested in behaviour [117]. The recent growth of interest in software agent technologies, e.g., [168], puts action at the centre of the stage. For software agents to have the capability of interacting intelligently with their environment they also need to be equipped with an ability to reason about what actions are the best for them to execute in given situations. In other words, intelligent agents need to be able to undertake practical reasoning. The most common response to this challenge has been to use some variant of the practical syllogism. In Section 2.1.3 of this chapter I will consider in detail the particular problems associated with the practical syllogism, first from the perspective of philosophy and then in Section 2.3.2, as seen in agent systems. However, before this I shall first consider the meaning and usage of the term ‘practical reasoning’, according to some existing literature.

Many definitions for term ‘practical reasoning’ exist from the large body of research on the topic. Without intending to discount the other definitions given by the many existing sources, I have chosen to examine a number of accounts of practical reasoning that closely relate to the issues explored in this thesis. I begin by going back to some of the earliest known literature on the topic, as documented by Aristotle.

In the collection of essays on Practical Reasoning in [138], Anscombe in her essay “*On Practical Reasoning*” uses the term *practical reasoning* synonymously with *practical syllogism* attributing the ‘discovery’ of this term to Aristotle. She discusses how this reasoning is commonly conducted in the same manner as theoretical reasoning - reasoning towards the truth of a proposition - which is supposedly shown to be true by the premises. The conclusion of such reasoning is of the general form ‘I ought to do such and such’. Anscombe gives Aristotle’s own example of the practical syllogism which appears in [7]:

Dry food suits any human
Such-and-such is a dry food
I am a human
This is a bit of such-and-such food

yielding the conclusion:

This food suits me.[138, p. 33]

Using this example Anscombe notes how Aristotle treats both theoretical and practical reasoning in the same manner by using syllogisms as a proof to necessitate the conclusion. In Section 2.1.3 I will examine the treatment of practical reasoning in this manner and discuss some of the problems associated with using this syllogistic form. Now, I turn a more recent definition of practical reasoning given by Bratman in [37]:

> "Practical reasoning is a matter of weighing conflicting considerations for and against competing options, where the relevant considerations are provided by what the agent desires/values/cares about and what the agent believes." [37, p. 17]

On this account, practical reasoning is highly dependant upon the situation the agent is in, and on its own individual desires. The important issue is to weigh up the competing options in order to try and determine the best one for the agent to execute at the particular time. The weighing up of such options is dealt with in this thesis through the use of argumentation techniques, as will be discussed in Section 2.4 of this chapter.

One of Bratman's major contributions to practical reasoning was the introduction of a new component to the traditional Aristotelian model of practical reasoning, which comprised solely beliefs and desires. This third component is the *intention*, which is a state of affairs that an agent has chosen to commit to bringing about [36]. This three component model – the BDI model – has proved to be extremely influential within the multi-agent systems community [168].

Bratman's definition of practical reasoning above also touches upon a particular point which is of great relevance in this thesis - the notion of *values*. Although Bratman does not place values in a category of their own, I argue in this thesis that values represent an important and distinct element of reasoning that needs to be accounted for in agent systems. Values are used in everyday human conversation so as to provide motivating reasons for having given aspirations. In this way people often refer to the set of 'values' that they hold. Such values can be wide ranging and could span anything from values held within a particular group or community, to more personal, individual values. In this sense values can direct human behaviour (either consciously or subconsciously) as part of the practical reasoning process, whereby people adopt goals that are intended to endorse the values held by the individual. Such value systems also provide us with an explanation as to why it is not always possible to persuade others to accept an opinion simply by demonstrating facts and proofs. It may well be that a particular individual will accept the facts of a particular decision but they may reject the conclusion to act because it does not support the values they hold. A discussion about such subjective disagreement based on the idea of subscription to different values is given by Searle in [146]. He addresses practical reasoning in his discussion of rationality and shows how and why disagreements occur in perfectly rational agents. Searle believes that the statement that rational agents should not disagree is an assumption about rationality which people often mistakenly make. In the introduction to his book "*Rationality in Action*" he states:

> "Assume universally valid and accepted standards of rationality, assume perfectly rational agents operating with perfect information, and you will find that rational disagreement will still occur; because, for example, the rational agents are likely to have different and inconsistent values and interests, each of which may be rationally acceptable." [146, p. xv]

This observation addresses the point of individual subjectivity and interpretation towards information and beliefs. Quite often in everyday life people have to 'agree to disagree' on matters because reasoned argument cannot resolve all conflicts. People can rationally subscribe to different beliefs and values which are in conflict with those of others. In the chapters that follow I argue that this point, in addition to other issues raised regarding rational argument in practical matters, should be applied to the practical reasoning mechanisms used in autonomous agents. The theory of persuasion over action which I am proposing incorporates the use of such values and I will show how this theory can be transformed into a computational agent setting which also makes use of such value functions.

The existing literature contains plenty of examples which discuss and make use of values, both within and outside a computational setting. One particularly prominent author on the subject is Perelman who believes that values have an important role to play in decision making, particularly in the justice system. To quote Perelman's book "*Justice, Law, and Argument*":

> "If men oppose each other concerning a decision to be taken, it is not because they commit some error of logic or calculation. They discuss apropos the applicable rule, the ends to be considered, the meaning to be given to values, the interpretation and characterisation of facts." [122, p. 150].

Thus, in Perelman's view it follows that it can be acceptable for one party in a discussion to disagree with another and this disagreement is accounted for through acceptance of differing value sets. Perelman's views in [122] refer in particular to the justice system as he goes on to say:

> "Each [party] refers in its argumentation to different values...the judge will allow himself to be guided, in his reasoning, by the spirit of the system i.e.,

> by the values which the legislative authority seeks to protect and advance."
> [122, p. 152]

Here, Perelman observes that judgments and justifications (of both legal and non-legal matters) made by a particular individual (or group) will depend upon the values held by the *listener*. Perelman refers to such an individual (or group) as 'an audience'. Thus, a particular audience may have a particular value preference which can differ from that of other audiences and so we can account for differences in opinion, even in matters where facts are agreed upon. The significance of this concept is that attention shifts away from the beliefs of the speaker and towards those of the audience. The notion of an audience has also recently been recognised in AI by Hunter in [84] and [85]. Although in these papers Hunter makes no distinction between reasoning about beliefs and reasoning about actions (though his examples do clearly involve reasoning about actions to be taken), like Perelman, he discusses the need to account for the fact that different audiences can have different perspectives on the same issue. He proposes an extension to a particular logic-based framework for argumentation that uses argument trees, which are a method that enable arguments and counter-arguments about a particular matter to be exhaustively collated. His extension to this framework is done through a model-theoretic evaluation of the believability of arguments. This in turn enables arguments to be ranked to have a more empathetic effect upon particular audiences. Such subjectivity in arguments with respect to a particular audience is obviously inherent in human reasoning. So, any account of practical reasoning for use in agent systems that is based upon human reasoning should also seek to model this subjectivity. I attempt to address this with respect to practical reasoning in the work I present in this thesis, although I do so in a different framework from that of Hunter, who's account does not stress the crucial distinction between arguing for a belief and arguing for an action. Perelman and Olbrechts-Tyteca discuss this notion of audiences and their preferences in further detail in *The New Rhetoric* [123]. I will also return to the topic later on in this chapter in Section 2.5.3 and again in Chapter 5 when discussing how audience preferences are captured through the use of Value-Based Argumentation Frameworks.

Values and audiences have featured prominently in other work from the legal domain. In [30] Bench-Capon and Sartor discuss in detail the role that values play in the legal justice system when reasoning about legal cases. Reasoning with cases has been viewed as a decision being deduced about a particular case through the application of a set of rules, given the facts of the case, e.g., [147]. However, the facts of cases are not set in stone as they can be open to interpretation from different lawyers. Additionally, the rules used to reach decisions are defeasible by their nature and as they are derived from precedent cases they too may be open to interpretation. In order to try and make the process of reasoning about cases more effective an alternative method can be used, which is the process of theory construction. This has been described by McCarty in [111]:

> "The task for a lawyer or a judge in a "hard case" is to construct a theory of the disputed rules that produces the desired legal results, and then to persuade the relevant audience that this theory is preferable to any theories offered by an opponent." [111, p. 285]

Thus, theory construction is intended to account for the context dependance of each particular case and the interpretation of the facts, rules and precedents involved. However, according to Bench-Capon and Sartor, this still leaves the problem of how to deal with conflicts amongst the rules that form the theory and how to choose preferences between these rules. The solution they propose is reached through the inclusion of values within the reasoning. Their idea stems from the work of Berman and Hafner [32] which suggests that the purposes of law need to be considered to explain why one particular rule is preferable to another. As the law is not composed arbitrarily, rather it is constructed to serve social ends, when conflicts in the application of rules occur in legal cases they can be resolved more effectively by considering the purposes of these rules and their relative applicability to the particular case in question. This enables preferences amongst purposes to be revealed, and then the argument can be presented appropriately to the audience through an appeal to the social values that the argument promotes or defends. In [30] Bench-Capon and Sartor support this idea with a formal model for theory construction and evaluation that takes into account the role that values play in the justification of the rules used in case-based reasoning. This method is again in line with Perelman's observations discussed above that audiences and values need to be accounted for in the presentation of argument.

Aside from the importance ascribed to values in law, there are plenty of other domains that also place importance on the notion. One particularly instructive case that has shown the significance of recognition of values is health promotion, as demonstrated by Grasso *et al.* in [73]. Here the authors address the issue of how to use dialogue to persuade people to change their diets to adopt more healthy eating practices. They account for the fact that conflicts can occur in people's opinions due to the different values and perspectives held by different parties. They use Perelman and Olbrechts-Tyteca's notion of *audiences* [123] to explain why some people may be more difficult than others to persuade into adopting healthy diets. The authors describe an agent, called Daphne, built upon a formalisation which allows dialectical argumentation regarding healthy nutrition to be conducted, in accordance with concepts from *The New Rhetoric*.

In addition to the domains I have briefly described above, numerous other domains could serve as examples of settings that place importance upon the use of values. The domains that will be used in this thesis to show the importance of modelling values in decision making are: politics, medicine, and law. In Chapters 6, 7 and 8 I will present specific example applications of decision problems that involve reasoning about values. For now, I examine the ways in which preferences have previously been represented in agent systems.

2.1.2 Modelling Preferences in Practical Reasoning

When applying practical reasoning to the design of autonomous computer systems we need some manner by which these variations in preferences can be expressed. In agent systems this has usually been done through the economic theoretic notion of utility functions ascribed to states. Utility functions represent the desirability of states, and agents act in order to try and maximize the perceived utility they expect to get from executing actions that lead to the most desirable states. However, this economic notion forces the assignment of rankings over states of affairs in a manner that seems counter-intuitive to the nature of practical reasoning. Returning to the work of Searle in [146], we can see that he holds the view that practical reasoning in humans does not occur through subscription to pre-existing utility functions:

> "This answer, [that an audience can provide a ranking for goals] though acceptable as far as it goes, mistakenly implies that the preferences are given prior to practical reasoning, whereas, it seems to me, they are typically the product of practical reasoning. And since ordered preferences are typically products of practical reason, they cannot be treated as its universal presupposition." [146, p. 253]

Thus, according to Searle, any theory of practical reasoning must take into account that choices concerned with the selection of actions should be made during the reasoning process and not form an input to it. The account given in this thesis aims at satisfying this criterion. The theory of persuasion over action presented in Chapter 3 makes use of values, in the sense of Perelman's account, and shows how preferences based upon individual values emerge through the practical reasoning process. Values, as used in the theory presented in this thesis, denote some actual descriptive social attitude/interest which an agent may or may not wish to uphold or subscribe to and they provide an actual subjective reason for wanting to bring about a particular state. In this sense values are not just a qualitative measure of a state, but they provide more subjective reasons as to why states of affairs are desirable or undesirable, even though values may themselves be qualified by labels that indicate their strength or importance. I discuss this point in more detail in Chapters 3 and 4 where I introduce a theory of persuasion over action incorporating the notion of values.

The observations, made above and also in Section 2.1.1, regarding the important features of practical reasoning have been recognised previously by computer scientists and some notable contributions have been given in recognition of these individual points. In the agents and argumentation literature, Fox and Parsons tackle a number of these points in [65] (which is an extension of their earlier paper on the subject [64]). In [65], Fox and Parsons recognise the differences immanent in reasoning about actions compared with reasoning about beliefs and discuss some of the pitfalls of standard decision theory [136] in dealing with reasoning about actions. Their main concern is the way in which a preference ordering on the expected utility of alternative actions is calculated. In standard decision theory this is done by assigning utilities to the outcomes of possible actions to produce the preference ordering, and assigning probabilities to the outcomes. However, as they point out, it is not any easy task to generate complete and reliable sets of probabilities and utilities for complex tasks, making quantitative representation impractical (as will be discussed further in Section 2.2 with reference to Searle's comments on the issue). Due to this deficiency, Fox and Parsons propose a qualitative approach to decision making in order to reduce the amount of numerical information required. To accomplish this they turn to the field of argumentation and in particular, they make use of a qualitative approach to the evaluation of arguments called the *Logic of Argumentation* (*LA*) [97]. LA was largely developed at the Advanced Computation Laboratory of the Imperial Cancer Research Fund UK with the motivation stemming from applications in the medical domain. Fox and Parson's work on LA is also discussed by Carbogim *et al.* in their review of argumentation [42]. In this review Carbogim *et al.* describe the main idea behind LA as being "to analyse the structure of arguments that are relevant to a particular proposition in order to obtain a degree of confidence for the proposition." This means that uncertainty in reasoning about beliefs is described in terms of arguments, rather than summative measures. Fox and Parsons make use of LA and propose an extension to it for dealing with reasoning about actions. Their extension proposes support for arguments about actions by incorporating the notion of expected values of actions. Their use of the word 'value' denotes a subjective assignment to a state (or condition) which produces a preference ordering on the states (and thus it differs from the sense in which the word 'value' is used elsewhere in this thesis where it is descriptive of a social attitude/interest). Their account [65] suggests that 'value assignments' (using their terminology) could be represented in a number of ways. The first of these 'value assignments' is through the adoption of a set of modal operators, such as *desirable(P)* or *undesirable(P)*, where *P* is some sentence such as "the patient is free of disease". A statement as to whether a proposition is desirable or not may provide us with a more qualitative assessment of an argument, but it does not tell us anything about the desirability of the argument in relation to other arguments in the debate. Alternatively they suggest labelling states with a sign, '+' to denote the positive valuation of a state and '-' to denote negative valuation of a state. However, using such an assessment we would also need some method by which we can determine which of the positively valued states is the most preferable. The account I present in Chapter 3 onwards is intended to address this issue. The final qualitative method for evaluating arguments proposed by Fox and Parsons is use of a dictionary of numbers to represent the possible monetary value of states. Based on this use of value assignment they describe how such arguments using values can be combined and they also discuss of how the expected value of outcomes of actions can be reasoned about and used to form preference orderings, to enable a decision to be made about which of the alternative actions to choose. The account provided in [65] can be seen as recognition of the need for subjectivity in practical reasoning. However, the account of practical reasoning that I provide in this thesis allows for a more fine-grained consideration of preferences with respect to audiences, as will be seen later on. The key point to note here about Fox and Parson's contribution is their recognition that classical decision theory cannot always provide us with reliable guidance as to what action to take. They also recognise that the preferences involved are solely based upon those of the agent making the decision. This ties in with Perelman's observations that different audiences have different perspectives, and Searle's comments that rational disagreement can occur due to different agents having different values and interests.

The accounts of practical reasoning discussed in this section are just a few taken from a large amount of research on the topic and they are intended to give a broad overview of the subject. The work presented in the forthcoming chapters of this thesis articulates an account of practical reasoning tailored for use by autonomous software agents. In particular, it takes the notions of values and audience preferences as concepts which can give valuable insight into the effective treatment and use of practical reasoning. Subsequently, I show how the use of these concepts in a computational setting enhances our ability to realise effective computational decision making in agent systems in a range of domains.

2.1.3 Difficulties with the Practical Syllogism

In this subsection I return to the practical syllogism and highlight some of the problems associated with it. As discussed in Section 2.1.1, practical reasoning within philosophy has been a topic of attention since at least the time of Aristotle. Recent discussions include collections of essays [115, 138] and a book by Searle [146]. Most of this work has taken as its starting point a version of the practical syllogism. Here is a typical example, taken from [92]:

K1 I'm to be in London at 4.15.
If I catch the 2.30 train, I'll be in London at 4.15.
So, I'll catch the 2.30 train.

Although Aristotle presented practical reasoning as a deduction, it has proved difficult to maintain that position (e.g., Anscombe's essays on the topic in [138] and Searle's remarks from [146] that will be discussed in Section 2.2) and this abductive form is now normally used. A problem is that it is possible to accept both of the premises yet deny the conclusion, based on at least three points of criticism:

C1: K1 represents a species of abduction, and so there may be alternative ways of achieving the goal.

C2: Performing one action typically excludes the performance of other actions, which might have other desirable results; these may be more desirable than the stated goal.

C3: Performing an action typically has a number of consequences in addition to the explicitly stated goal. If some of these are undesirable, they may be sufficiently bad to lead us to abandon the goal.

In order to act on the basis of an argument such as K1, therefore, we need to consider alternative actions, alternative goals and any additional consequences, and then choose the *best* of these alternative goals and actions. Note the element of choice here: we can choose which of our goals we will seek to realise, and which actions to undertake to realise these goals. This freedom is the essence of "autonomy". Such an element of choice regarding the adoption of goals differs somewhat from the choice associated with beliefs. In situations where we know what is or what is not the case we do not have a choice in the matter about what we believe. No rational agent can believe the opposite of what it knows. However, in the absence of complete knowledge we can make choices as to whether or not to believe facts told to us by others. Moreover rational agents can be compelled - so long as they remain rational - to believe or disbelieve information given to them through demonstration and explanation on the part of others. Such coercion also involves influences from social factors such as trust, motive and reliability of the source, and these too play a large part in whether or not we choose to believe what others tell us. Furthermore, in the light of new and updated information we may also choose to revise our beliefs and adopt or disregard new information presented to us [1]. However, the choices associated with the adoption of goals and actions to perform yield more flexibility as they are not based upon proofs as to what is or is not the case. Unlike beliefs as to what is true, when the world being as it is means that there is a right and wrong answer, different people may rationally make different choices of goals and actions. We are not driven by our desires: we can resist them. And whereas the way the world is lies beyond our control, we can at least (to some extent) choose the way we would like the world to be.

Given this element of choice therefore, practical argument is directed to a specific person at a specific time, to encourage them towards a particular choice of goals and/or actions; the objectivity that we can find in factual matters cannot in general be attained in practical reasoning. An attempt to modify K1, similar to one put forward by Searle in [146] (although not regarded by him as satisfactory) is:

S1 I want, all things considered, to achieve E
The best way, all things considered, to achieve E is to do M
So, I will do M.

The two different “all things considered” qualifications are supposed to deal with alternative desires and methods of achieving them. The “best” addresses the selection from the available options. However, this too presents problems: we cannot in general consider all things, because we have limited reasoning resources and imperfect information. Nor is it easy to say what is meant by “best” here. In computer science there are often attempts to define best using some kind of utility function but, as discussed in the previous subsection, Searle points out that any preference ordering is more often the *product* of practical reasoning than an input to it. Coming to understand what we think is best is part of what we do in practical reasoning. The syllogistic treatment of practical reasoning does not account for such subjective differences. These issues become even more relevant when dealing with the construction of reasoning mechanisms in agent systems.

In the next section I will explore Searle’s work on the nature of practical reasoning in more detail and in particular I will survey his discussions on the non-deductive nature of practical reasoning. His work on the topic provides a contribution which helps to explain why the difficulties with the practical syllogism, as discussed in this subsection, do in fact occur. Additionally, Searle’s account of practical reasoning will be the one I shall follow in the rest of this thesis.

2.2 Searle’s Account of Rationality in Action

In the previous section I noted a number of interesting points raised by Searle regarding the nature of practical reasoning. In this section I examine in more detail the account of practical reasoning given by him in his book “*Rationality in Action*” [146]. In this book he makes a number of objections to the “classical” model of rationality and proposes some solutions to overcome the weaknesses that he identifies in the model. A number of observations that he makes are instrumental to the theory of persuasion over action that I present in Chapter 3 and my theory aims to take some of these important observations into account. I begin this discussion by recapitulating Searle’s objections to the classical model of rationality.

2.2.1 Objections to the Classical Model of Rationality

Searle begins his discussion of rationality in action by stating six assumptions that lie behind the classical model of rationality and then he presents his doubts in relation to each assumption. Before I summarise these assumptions and associated doubts I will first clarify what Searle means when he refers to the "Classical Modal of Rationality." He acknowledges that there is no unifying definition of the model and that many philosophers, such as Aristotle, Hume and Kant do not share exactly the same conceptions of rationality. However, following Hume's claim that "reason is, and ought only to be, the slave of passions" [83], Searle believes that Hume gives the clearest statement of what Searle refers to as the classical model. This model has the underlying notion that reasoning about action is driven by desires, and the most sophisticated manifestation we have of this model is presented in contemporary mathematical decision theory. Searle begins his discussion by providing a motivating example to explain how he first came to consider the problems with mathematical decision theory. I shall now give a brief reprisal of this example.

Consider the situation where I value my life and I value twenty five pence. It would then seem to be a strict consequence of the axioms of decision theory that there must be some odds at which I would bet my life against twenty five pence. Searle considered this and came to the conclusion that there are in fact no odds at which he would bet his life against twenty five pence, and even if there were, he would not bet his child's life against twenty five pence. A number of decision theorists challenged Searle's conclusion on this matter and they themselves concluded that his thinking here was 'plain irrational'. Searle rejects this claim believing that the problem lies not in his way of thinking, but in the limitations of this decision theoretic model of rationality. In an attempt to explain what he believes to be the limitations of this classical model of rationality Searle presents a list of six assumptions that lie behind this classical model and he uses them to express some doubts he has about the model. I summarise these assumptions and doubts below.

1. **Actions, where rational are caused by beliefs and desires.**

   Searle's criticism of this point is that he believes it to be incorrect to assume that a person's set of beliefs and desires is *causally sufficient* to determine an action. On the contrary, he states that in a typical case of rational decision making, a person has a choice of alternatives available and considers the various reasons for choosing one of the options. This stems from the human attribute of "freedom of will". Searle refers to free will as 'the gap', which exists between the causes of action in the form of beliefs and desires and the effect in the form of the action.

2. **Rationality is a matter of obeying rules, the special rules that make the distinction between rational and irrational thought and behaviour.**

The doubt Searle expresses about this point is that rationality is a broader concept than just following rules[1]. This issue is again related to the existence of ‘the gap’ whereby humans have the same gap for inferring as we do for any other voluntary activity. Searle states that “we also need to distinguish between entailment and validity as logical relations on the one hand, and inferring as a voluntary activity on the other.” Inferences are valid in cases where the premises entail the conclusion, but there is nothing that actually forces a person to make that inference. Thus, in real life reasoning we must choose to believe the semantic content of inferences, as validity of the inference cannot be guaranteed solely through the application of a syntactic rule. Even if we are reluctant to concede this in matters of fact, Searle seems right with respect to moral choices.

3. **Rationality is a separate cognitive faculty.**

   Searle criticises this statement by pointing out that rationality cannot be separated from the human capacities such as language, thought, perception and various forms of intentionality.

4. **Apparent cases of weakness of will can only arise in cases where there is something wrong with the psychological antecedents of the action.**

   Here Searle argues that weakness of will is always possible no matter how perfectly you structure the antecedents of your action. This arises again due to the presence of ‘the gap’, which at any point provides an indefinitely large range of choices open to people and some of these choices will seem attractive, even if a decision has already been taken to refuse them.

5. **Practical reason has to start with an inventory of the agent’s primary ends, including the agent’s goals and fundamental desires, objectives and purposes; and these are not themselves subject to rational constraints.**

   In criticising this point Searle argues that we can act from obligation as well as desire. If we restate actions motivated by recognition of an obligation as motivated by a desire to meet the obligation, we lose a useful distinction.[2]

6. **The whole system of rationality works only if the set of primary desires is consistent.**

   The point Searle makes here is that it is common for humans to have inconsistent ends. We may desire a number of things, whilst at the same time being aware

[1] This point is also endorsed by Shackle in [148] who argues that a rational person would not blindly follow an algorithm for decision-making regardless of circumstances, so rationality cannot be defined by adherence to a rule.

[2] Other philosophers, such as Frankfurt [66], have argued that desires are themselves produced by a rational process, involving the formulation of “second-order desires” e.g., I have a desire that I desire to do more exercise.

that we cannot achieve all of them at once. The task of rationality in practical reasoning is to try to find some way to adjudicate between various inconsistent aims.

The above list is intended to give a brief summary of Searle's criticisms of the assumptions that lie behind the classical model of rationality. He believes that due to these incorrect assumptions, mathematical decision theory does not give a general theory of the role of rationality in thought and action. In [146] Searle gives further in-depth discussions of each of the above points. All these criticisms contribute towards a larger issue that Searle deals with towards the end of the book: why there is no deductive logic of practical reason. I shall now briefly summarise the points that Searle makes regarding this matter.

2.2.2 Why There is No Deductive Logic of Practical Reason

In considering the issue of the non-existence of a formal logic of practical reason Searle naturally considers why and how there is a deductive logic of theoretical reason. He begins this discussion by giving a brief statement as to why we have a deductive logic for theoretical reasoning. Following Frege's work on deductive logic [67], we can distinguish between two separate aspects of theoretical reasoning: "logical" notions of premise, conclusion and logical inference, and, "psychological" notions of belief, commitment and inference. Searle points out that there is a "tight set of parallels" between the two notions as the features of logical consequence can be mapped on to the commitments of belief, because logical consequence is truth-preserving. Thus, if *q* is a logical consequence of *p*, and I believe *p*, then I am committed to the truth of *q*. Following this principle, that belief in the premises of a valid argument in theoretical reasoning commits you to belief in the conclusion, Searle wishes to investigate whether we can get similar commitments to desires and intentions as conclusions in practical reasoning. He articulates this point by posing the following question:

> "Are there formal patterns of practical validity, such that the *acceptance* of the premises of a valid practical argument commits one to the acceptance of the conclusion, in the way that is characteristic of theoretical reason?" [146, p. 241].

Searle begins discussion of this question by considering means-end reasoning. If a person is committed to achieving a particular ends, such as going to London, then he must try to find a means by which this can be achieved, such as catching a train to London. However, it may well be the case that desiring the ends does not in fact commit one to desiring the means. For example, in wanting to go to London, there may be a number of other means by which this end can be achieved, such as catching a plane, walking, getting a boat, etc., and one of these means may be more desirable to commit to. This shows that entailment relations do not generate commitment to a secondary desire in practical reasoning. To put this more precisely and in Searle's words: "if I believe both *p* and 'if *p* then *q*' then I am committed to the belief that *q*. But, if I want *p* and I believe that 'if *p* then *q*', I am not committed to wanting *q*". This points out a clear difference between practical and theoretical reasoning in terms of entailment relations and Searle goes on to consider why this difference exists.

In Section 2.1.3 I provided a short example detailing Searle's attempt to modify the practical syllogism to overcome the problem of commitment to means as well as ends, as discussed above. Searle's modification of the practical syllogism includes the qualification in the premise that in seeking a means to an end one must consider "all things" in order to be able to find the "best" means for achieving the particular end. This qualification enables us then to discount any means that we do not consider to be the "best". However, he does not deem the addition of this qualification to solve the problem of logical relations in practical reasoning. This is because it is not clear how we can actually give a generally applicable definition of the term 'best' and what consideration of 'all things' actually constitutes. Additionally, he also points out that there is no need for an equivalent qualification to be added to the premises in theoretical reasoning and therefore this proposal does not in fact map the logical relations on to the psychology in the correct manner. So, in an attempt to understand why a logic of practical reasoning cannot be constructed in a manner analogous to the logic of theoretical reasoning, Searle turns his attention to exploring the structure of desires and the differences between beliefs and desires.

The first difference that Searle considers here is that it is possible for an agent to consistently and knowingly want that *p* and also want that *not p*, in a way in which it is not possible for an agent to consistently and knowingly believe that *p* and *not p* are both the case. For example, I may have two separate desires which involve me wanting to be in two places at the same time, but knowing that these desires are inconsistent I cannot rationally commit to wanting the conjunction of both desires. It is a logical consequence that desire is not closed under conjunction: if I desire that *p* and I desire that *not p*, then it does not follow that '*I desire p and not p*'. Searle attributes this anomaly to the existence of primary and secondary desires. To use one of his examples, suppose I have a desire to buy a plane ticket. However, I do not want to buy a plane ticket for the sole purpose of possessing such a ticket, but rather my desire is motivated by some other desire, e.g., that I want to go to London. Here the primary desire is wanting to go to London and the secondary desire is wanting to go by plane. However, it may be possible for an agent to formulate a set of conflicting secondary desires, given a set of primary desires and beliefs about the best means of satisfying them. Returning to the example, as well as my having a desire to go to London, I may also have a desire not to travel anywhere by plane, in order to satisfy another primary desire of not participating in activities that invoke my fear of flying. This desire of not travelling by plane is clearly in conflict with my original desire which states the opposite. Searle summarises this point by stating:

> "the same person, using two independent chains of practical reason, can rationally form inconsistent secondary desires from a consistent set of his actual beliefs and a consistent set of primary desires." [146, p. 252].

The solution to the problem of having such inconsistent desires, and a need to choose only one to commit to, manifests itself in the notion of *preference*. Again, returning to the travel example, although I may have two inconsistent desires of wanting and not wanting to travel by plane I must choose one to commit to by expressing my preference. It may be the case that I prefer to take the flight and endure my fear of flying, or vice versa, and I will choose which desire to commit to, according to the option I prefer. However, as I noted in Section 2.1.2, Searle believes that the proposed solution of appeal to preference ordering mistakenly implies that preferences form input, as opposed to output, of the practical reasoning process. Given this he concludes that:

> "...even if we confine our discussion of practical reasoning to means-ends cases, it turns out that practical reason essentially involves the adjudication of conflicting desires and other sorts of conflicting motivations [...] in a way that theoretical reasoning does not essentially involve the adjudication of conflicting beliefs." [146, p. 253].

Additionally, this conclusion brings forth further problems in that:

> "The Classical conception works on the correct principle that any means to a desirable end is desirable at least to the extent that it does lead to the end. But the problem is that in real life any means may be and generally will be undesirable on all sorts of other grounds, and the model has no way of showing how these conflicts are adjudicated." [146, p. 254].

Thus, Searle believes that practical reasoning in itself does and should typically involve the adjudication of conflicting desires, needs, commitments etc. But, the classical model is flawed in this provision through the absence of any mechanism by which we can decide what constitutes the 'best' way to do something and how we can reconcile inconsistent conclusions of valid derivations.

In recognition of Searle's conclusions on this matter, the theory of persuasive argument in practical reasoning that I present in this thesis attempts, to a certain extent, to address such problems. The theory presented in the following chapters aims at providing a model that accounts for conflicting preferences in practical reasoning and also provides a method whereby conflicts can be rationally reasoned about and adjudicated to some acceptable degree.

Searle summarises his discussion on the nature of desire by defining two special features of desire that make it impossible to construct a formal logic of practical reasoning that is analogous to our formal logic of theoretical reasoning. The first of these features Searle labels "the necessity of inconsistency" whereby any real-life rational agent is bound to have inconsistent desires and motivators. The second feature he labels "the non-detachability of desire" whereby sets of beliefs and desires that form the premises of the practical reasoning do not necessarily commit the agent to having a corresponding desire as a conclusion, even where the propositional content of the premises entail the propositional content of the conclusion: modus ponens does not produce a commitment to desiring a conclusion from desire/belief combinations.

Searle then turns his attention to focusing on a explanation as to why desires differ from beliefs. He believes this can be attributed to a particular underlying feature of desires: "that they are inclinations towards states of affairs (possible, actual or impossible) under aspects" and there is no commitment involved here. Opposingly, beliefs are convictions that states of affairs exist under aspects and the agent is committed to such beliefs. It is not possible to rationally accept that a state of affairs both does and does not exist under the same aspect, but it is possible to be both rationally inclined and disinclined to the same state of affairs under the same aspect. It is the commitment element that distinguishes the two: commitment to actual states of affairs excludes the possibility of rationally inconsistent beliefs, whereas there is no commitment involved in the way the agent would like the world to be, and therefore there is no restriction on consistency of desires and no logical consequence can necessarily follow. However, the case is different when we turn our attention to intentions. Searle gives a brief discussion of this, which is his final point in addressing the question of why there is no deductive logic of practical reasoning. I will now summarise his remarks on intentions.

Like beliefs, intentions have the possibility of being inconsistent. The role of an intention is to cause some action to be performed and intentions are causally self-referential in that an agent's intention is only carried out if the agent acts by way of carrying out the intention. Thus, it is not possible to hold two inconsistent intentions, as inconsistent actions cannot both be carried out. Additionally, and again like beliefs, intentions also incorporate the notion of commitment. Intentions involve commitment to the satisfaction of the intention by way of action. However, the important thing that Searle notes here is that an agent is:

> "not committed to intending to achieve all of the consequences of the achievement of his intention. He is committed only to those means that are necessarily intended in order to achieve his ends." [146, p. 267].

This observation accounts for the issue of side effects of the execution of actions. It may be the case that a specific action has side effects that bring about states of affairs that the agent does not will. An intention to bring about *p* and a belief that *if p then* *q* does not commit an agent to having an intention to bring about *q*. An example of this, adapted from one of Searle's examples [146], can be seen in medical treatment. If a person is ill they may have to undergo treatment prescribed by a doctor to cure them of their ailment, but an undesirable side effect of this may be that the person has to suffer pain as part of the treatment. It is not the doctor's original intention to cause pain, as the original intention is to cure the ailment. This point can be further qualified by assessing the success of bringing about the intention. If in fact no pain is caused by the treatment then the belief that this would happen has turned out to be false, but the intention has not failed (as long as the patient is actually cured of the ailment). Searle raises this particular point in objection to Kant's famous claim that "he who wills the ends wills the means" [90]. Although Searle does not believe that this statement is false in all cases, he does believe that it does not hold in all cases of practical reasoning. To summarise the point Searle states:

> "it is simply not the case in general that anybody who wills the end (in the sense of having an intention to achieve that end) thereby wills everything that occurs as a known part of carrying out that intention." [146, p. 265]

As will be seen in the chapters to come, the existence of side effects of actions plays an important role in the assessment of the suitability of actions in my account of persuasion in practical reasoning.

This final point concludes Searle's presentation of reasons and explanations as to why there cannot be a deductive logic of practical reasoning, in the sense that there exists such an account for theoretical reasoning. In the next section I will turn my discussion from practical reasoning in philosophy to practical reasoning in autonomous software agents, and in particular in agents based on the Belief-Desire-Intention architecture. The discussion will take forward some of Searle's arguments, as presented in this section, to explain how the application of the practical syllogism in autonomous agents also raises problematic issues.

2.3 Autonomous Agents and the Belief-Desire-Intention Architecture

In this section I examine how practical reasoning is generally deployed in autonomous agents based on the popular Belief-Desire-Intention architecture. Traditionally, this has been done through the application of the practical syllogism to an agent's reasoning mechanism. I discuss some of the shortcomings of this approach, which I will attempt to address in the theory presented in subsequent chapters.

2.3.1 Practical Reasoning in BDI Agents

The development and use of autonomous software agents is now a firmly established discipline within computer science. There is a wealth of literature in the area dealing with epistemic reasoning and logics (a standard textbook is [113]), which are essential cornerstones upon which agents are built. However, although it is widely recognised that reasoning about action is an essential activity for an agent to be able to perform, as mentioned in the previous sections, the treatment of practical reasoning in multi-agent systems has not received as much attention as reasoning about beliefs. The process of practical reasoning in computer science is often accomplished through the application of some form of the practical syllogism. Here I discuss some of the problems associated with the use of the practical syllogism in methods used by agents to perform practical reasoning. In particular, I discuss this in the context of the Belief-Desire-Intention architecture which is one of the most influential agent architectures discussed and deployed in agent systems. Because the BDI model has a number of proposed realisations, I will, when I need to be specific, take as my model the popular Procedural Reasoning System (PRS) [68], as this is widely used. The PRS system is depicted in the diagram below, as taken from Wooldridge's book "*An Introduction to Multi-Agent Systems*" [169, p. 83].

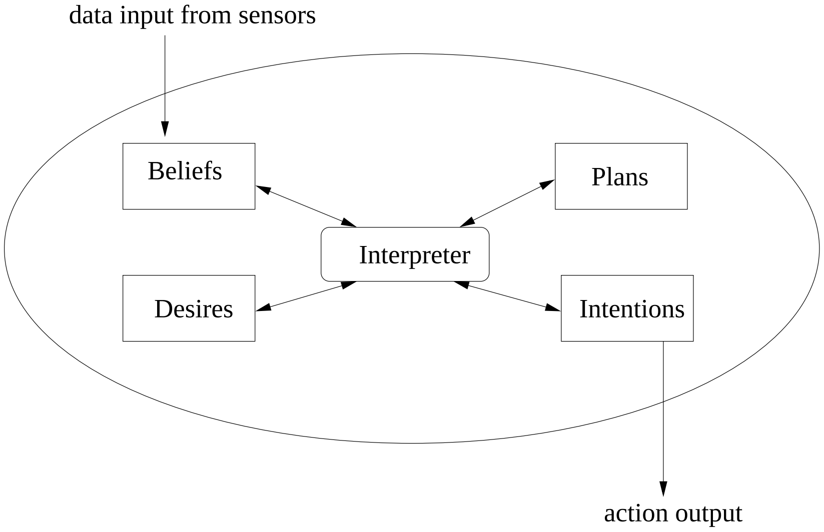


Figure 2.1: The Procedural Reasoning System (PRS) [169].

The process of reasoning about action is described by Wooldridge in [168] as "the Deliberation Process" and this process comprises two phases: option generation and filtering. During the option generation phase the decision-making agent generates a set of possible alternative actions available for execution. These alternative options are generated by taking the agent's current beliefs and current intentions and applying the reasoning scheme of PRS to see which options can now be pursued. These options form the current set of desires of the agent. Thus, the agent's desires correspond to the goals that it wishes to realise, though it may be the case that not all desires are achievable. In order to achieve these desires the agent must form a plan from the repertoire it holds in a pre-programmed plan library and check that the pre-conditions for executing this plan are satisfied by the agent's current beliefs about the world. This results in the agent developing a set of actions (or plans) in order for it to achieve its desires. The agent can now move on to the filtering phase where it simply chooses the "best" option to commit to from this set through the use of a filter function. The "best" option will typically be chosen through the application of some pre-existing utility function, and then added to the intentions of the agent.

2.3.2 Limitations of the Practical Syllogism in Computer Science and Agent Systems

Searle's form of the practical syllogism, as discussed in Section 2.1.3, can be applied to the reasoning mechanisms used in autonomous agents in order to equip them with the ability to reason about what it is best to do in a given situation. The standard view of the justification of an action in this context can be generally seen as:

PS1 Agent P wishes to realise goal G
If P performs action A, G will be realised
Therefore, P should perform A.

This view represents the option generation stage of deliberative reasoning in the Belief-Desire-Intention (BDI) model [168].

Thus, an agent using the BDI model is able to address some of the difficulties associated with the practical syllogism highlighted in Section 2.1.3 in the following ways:

- The agent has a repertoire of plans held in a finite plan library and this enables it to consider everything available to it that is relevant to the decision.
- The agent is able to define which action is the *best* one to take as it has a utility function, or some other filtering criterion, to enable it to compare potential outcomes of actions.
- Any undesirable side effects, brought about as a consequence of performing an action, cannot be considered as ruling out the plan as the agent's plan library

> should contain only plans approved by the designer, recognising these side effects. The agent should be able to assume that it is permitted to carry out any plan it has been given.

While this approach provides a pragmatic resolution of the issues appropriate to some agent systems, it provides a less satisfactory solution to the general problems associated with practical reasoning as discussed in Section 2.1. By its nature, the process of practical reasoning is open-ended and this in turn poses problems for its use in agent technology. Agents operate with a limited repertoire of plans and a fixed utility function and so the designer necessarily takes responsibility for pre-determining the options available to the agent. The agent can consider only the options it has been given, not "all things". Even with the autonomy afforded to agents, constraints are made upon the plans the agent will consider and find acceptable, as filtering of alternative plans is undertaken by means of a fixed utility function over goals, supplied in advance by the designer[3]. Because practical reasoning is intrinsically open-ended, unforeseen alternatives and consequences may arise, and revision of preferences may occur, at any time. This creates a challenge for agent design, which must, by its nature, make assumptions which circumscribe the considerations possible to the agent. These problems become even further exacerbated when designing agents to be part of a collective with group objectives and values. In the work presented in the coming chapters I attempt to address some of the problems highlighted here. I do so by turning to the field of argumentation theory and in particular by accounting for practical reasoning in terms of argument schemes and critical questions, as will be discussed in Section 2.4.

Before I discuss the topic of argumentation theory I shall first examine one particular model of practical reasoning that has been computationally implemented upon an extension of the BDI model. This is the OSCAR project of Pollock [127]. The objectives of the project were twofold: to construct a general theory of rational cognition, and to construct an artificial rational agent to implement this theory. OSCAR addresses these objectives within an architecture that distinguishes between theoretical reasoning (what Pollock calls 'epistemic cognition') and practical reasoning (what Pollock calls 'practical cognition'), providing a model for both types of cognition, with the two aspects interacting. The overall architecture is built upon a defeasible reasoner, which enables OSCAR to defeasibly deal with perception, change and persistence, causation, probabilities and plan construction and evaluation.

The model of practical reasoning used in OSCAR is based on Pollock's Belief-Desire-Intention-Liking (BDIL) model. This extends Bratman's BDI model of prac-

[3] It is in this way that agents differ from humans: agents cannot themselves desire to change their desires. The ability of humans to form desires about what they desire is what distinguishes our level of autonomy (what we call freedom of will) from that of a software agent. The formulation of desires is an interesting issue that has implications for autonomy in multi-agent systems. As mentioned in Section 2.2, one interesting discussion of human freedom of will has been given by Frankfurt [66] who attributes our free will to our ability to form "second-order desires".

tical reasoning [36] by adding a new component to the model, which Pollock calls ‘likings’. These ‘likings’ are split into two types: ‘situation likings’, which are states of the world that the agent likes, and, ‘feature likings’, which are particular aspects of a situation that an agent likes [126]. These ‘likings’ are also supplemented with three types of desires, in addition to standard beliefs and intentions. This results in a seven component model of practical reasoning.

According to Pollock, the OSCAR architecture differs from most agent architectures in that, even though the agent’s interactions with the world are directed by practical reasoning, most of the work that forms the rational cognition is performed through theoretical reasoning. This is done through the following process:

- Practical cognition evaluates the world (as represented by the agent’s beliefs), and then poses queries concerning how to make it better.
- These queries are passed to epistemic cognition, which tries to answer them.
- Competing plans are evaluated and selected on the basis of their expected utilities, but those expected utilities are again computed by epistemic cognition.
- Finally, plan execution generally requires a certain amount of monitoring to verify that things are going as planned, and that monitoring is again carried out by epistemic cognition.

Pollock summarises the approach by stating that in general, the choices made in OSCAR are decided upon by the practical cognition component, but the information on which the choices are based is the product of the epistemic cognition component, with the majority of the work in rational cognition going into providing that information. A description of the general architecture and theory upon which OSCAR is built is described in detail in [126].

A nice overview of OSCAR and its merits has been given by Hitchcock in [82]. In that paper Hitchcock also gives a summary of what he believes to be the main weakness of the OSCAR model, which I will summarise briefly here. Firstly, Hitchcock believes Pollock’s model to be solipsistic, in the sense that there is no mechanism for communication between agents and thus discussion and/or exchange of arguments about an issue is not an option. Secondly, the model is egoistic, in the sense that the purpose of the agent is to tailor the world to its own liking as far as possible. This means that the agent has no regard for the likings of other such systems in its environment. Finally, the model is unsocial, in the sense that OSCAR does not, and cannot, belong to any group of autonomous rational agents with decision making capabilities regarding the group’s action. Hitchcock concludes that any model of practical reasoning needs to address these deficiencies. My model for practical reasoning, detailed in the coming chapters, overcomes some of the general shortfalls that are associated with OSCAR, as highlighted here.

This concludes my discussion of practical reasoning in autonomous agents. I will now examine the field of argumentation theory and the methods it presents for dealing with uncertain information, which is an inherent feature of multi-agent systems.

2.4 Argumentation Theory

In this section I discuss the philosophical background of the representation and treatment of argument in philosophy. In particular I examine one account of argument representation based upon the use of argument schemes, which is shown in this thesis to be of benefit to the application of practical reasoning in multi-agent systems.

2.4.1 Argument Schemes

The study of argumentation theory within the field of philosophy has a rich body of literature dating back to the time of the ancient Greek Philosophers. The use of argument has proved to be extremely useful in contexts where proof cannot be used, e.g., in domains where information is incomplete, uncertain or implicit. An argument is less tightly specified than a proof, as arguments offer open-ended defeasibility whereby new information can be brought to an issue and the reasoning can proceed non-monotonically. The same is not true of a proof, as I discussed in Section 2.2 with regard to Searle's analysis of a formal deductive logic of theoretical reasoning. Also, arguments provide us with a concept of subjectivity that is not present within a proof: an audience can choose to reject an argument, whereas they cannot reject the conclusion of a proof, if they accept the premises, and if they accept the rules of inference. Proofs play an essential role in matters where information is certain and complete, e.g., in Mathematics. However, most real-world situations do not have such clear cut information available and it is here where argument plays its important role. This point was emphasised by Perelman and Olbrechts-Tyteca in [123] who state:

> "Logic underwent a brilliant development during the last century when, abandoning the old formulas, it set out to analyze the methods of proof used effectively by mathematicians. [...] One result of this development is to limit its domain, since everything ignored by mathematicians is foreign to it. Logicians owe it to themselves to complete the theory of demonstration obtained in this way by a theory of argumentation." [123, p. 10].

According to Perelman and Olbrechts-Tyteca, argumentation has the potential to provide us with a means to complement mathematics by addressing the issues that cannot be solved by mathematics alone.

There are many potentially useful ways to approach argument representation. The approach that is used in this thesis is based upon the use of presumptive reasoning and argument schemes. By this method the arguments are presented as general inference rules whereby given a set of premises, a conclusion can be drawn. However, such schemes are not deductively strict due to the defeasible nature of arguments. The schemes allow for arguments to be represented within a particular context and take into account that the reasoning presented may be altered in the light of new evidence or exception to rules. Such schemes have proved to be of benefit in a number of areas including AI, law, informal logic and the study of fallacies.

One early example of the use of argument schemes is *Toulmin's Argument Schema* [157]. One of the main features of this schema is that in contrast to previous schemes for argument that have been based upon logical proofs consisting of the traditional premises and conclusion, Toulmin's allows for more expressive arguments to be asserted. It does so through the incorporation of additional elements to describe the different roles that premises can play in an argument. Toulmin's schema comprises the following elements:

- a *claim*, which is the conclusion of the argument;
- a *qualifier*, which gives the strength of the argument for the claim;
- the *data*, which is like a traditional premise;
- the *warrant*, which licences the derivation of the claim from the data;
- a *rebuttal*, which is a proposition which would refute the claim, if the rebuttal were to be proved true;
- the *backing*, which represents the authority for the warrant.

The elements of this schema are connected as shown in Figure 2.2.

Toulmin's schema proved to be popular due to the expressivity it afforded in the presentation and justification of arguments. It has been the focus for a number of implemented systems [31, 103, 173] due to its novel structure enabling the defeasible nature of arguments to be reasoned about more effectively than previous logic based schemes. One such implementation in the form of a dialogue game has been undertaken in [24]. The application used in that particular example is to enable the conduct of effective legal reasoning. However, although the schema has proved to be an influential contribution, it lacks elements that have proven to be of use in dealing with the precise identification of conflicts in arguments. For example, unlike the critical questions associated with certain argument schemes (e.g., Walton's schemes in [164], which are subsequent to Toulmin's) Toulmin's schema says little about the manner in which

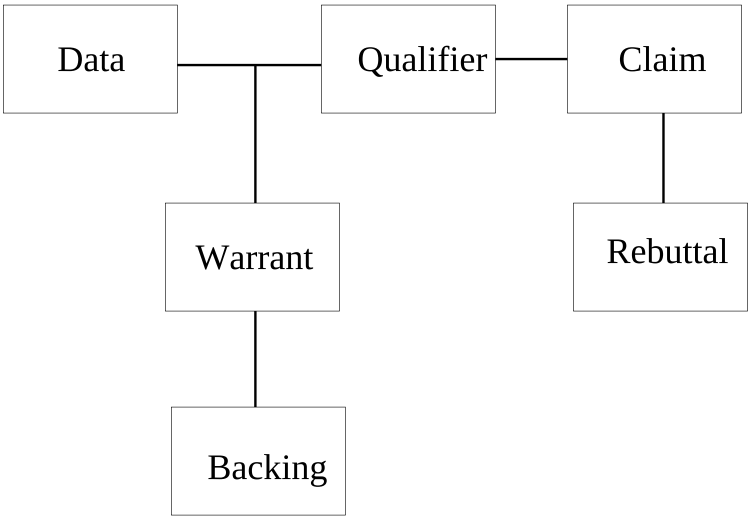


Figure 2.2: Toulmin’s Argument Schema [157].

the argument can be attacked. Although the schema does take into account that the claim could be challenged through the use of the rebuttal, it does not provide a detailed manner in which an opponent can explicitly attack elements of the argument. Indeed, opponents of an argument are intended to question and request further information to be supplied in the form of the data, warrant, backing and qualifiers. However, using this schema there is no way to distinguish between different kinds of attacks, such as a rebutter (which is a proposition that refutes the claim) or an undercutter (which refutes the inference made between the premise and conclusion [126]), so the precise nature of the disagreement may not always be easy to identify. Also the notion of values plays no separate role in Toulmin’s schema. It is possible to imply value based arguments through the statement of a warrant in Toulmin’s schema. However, the warrant may equally involve a statement of facts or beliefs about the world, making no categorical distinction between the values and beliefs. As highlighted earlier, the theory presented in this thesis places importance upon a distinct notion of value and thus Toulmin’s schema is unable to state precisely the expression of arguments in such terms. In addition to this, Toulmin’s schema makes no separation of arguments regarding beliefs and those regarding actions. As discussed earlier in Sections 2.1 and 2.2, these two forms of reasoning have a distinct nature and contrasting objectives for the outcome of the reasoning process. Thus any account of practical reasoning needs to be dealt with in a separate manner from reasoning about beliefs. Toulmin’s schema is recognised as an influential contribution to the field, but due to the issues highlighted above I believe that there are more precise and expressive ways to structure arguments about actions, as I will demonstrate in the following chapters.

Since the introduction of Toulmin's schema, great advances have been made in dealing with the representation of arguments in terms of schemes. One significant contribution to the field has been Walton's notion of argument schemes and associated critical questions, as discussed in more detail in Section 2.4.2.

The concept of presenting argumentative reasoning in the form of argument schemes has proved to be a fruitful method for classifying different kinds of arguments, in order to deal with each in an appropriate fashion. In the legal domain Verheij has shown in [161] that argument schemes can be embedded in a particular formal dialectic logic in order to be deployed in legal reasoning. A further example of the application to the legal domain has being given by Prakken *et al.* in [130], showing how argument schemes can be used to reason about evidence. The same authors have also examined in [131] how the burden of proof can shift during persuasion dialogues using one of Walton's particular argument schemes – the argument from expert opinion [165].

Looking towards the implementation of argument schemes, a notable example is the argument visualisation software *Araucaria* of Reed and Rowe [140]. *Araucaria* is a system designed to provide support for the analysis and diagramming of arguments. It uses the Argument Markup Language (AML) (an open standard, designed in the Extensible Markup Language, (XML)), to describe the structure of arguments. It has been used in a number of illustrative examples such as [131] and [141] to show how arguments can be formally represented and reasoned about using the tool. *Araucaria* however, is a tool for visualising and manually constructing arguments and it offers no automated support for reasoning about arguments.

In the next two subsections I examine one particular argument scheme more closely - Walton's scheme for practical reasoning - and describe how this can embody a particular setting of practical reasoning known as presumptive reasoning.

2.4.2 Presumptive Reasoning and Argument Schemes

I now examine how one particular notion of practical reasoning – presumptive reasoning – has been treated through the use of argument schemes.

One way of addressing the problems associated with the practical syllogism, highlighted in Section 2.1.3 of this chapter, is to regard practical reasoning as a species of presumptive argument: given an argument, we have a presumptive reason for performing the action. This presumption can, however, be challenged and withdrawn. Subjecting our argument to appropriate challenges is how we hope to identify and consider the alternatives that require consideration, and determine the best choice for us, in the particular context. Because the challenges are, in principle open ended, the process of justification does not end, and discussion can always be re-opened.

One account of presumptive reasoning is in terms of argument schemes and critical questions, in the manner discussed in the previous subsection and as proposed by

Walton in [164]. In this account an argument scheme is viewed as embodying a presumption in favour of the conclusion. Whether this presumption stands or falls can be tested through posing critical questions associated with the scheme. In order for the presumption to stand, satisfactory answers must be given to any such questions that are posed in the given situation. I discuss this particular approach in further detail below.

2.4.3 Walton's Account of Practical Reasoning

In [164] Walton gives two schemes for practical reasoning: the *necessary condition scheme* (called W1):[4]

W1 G is a goal for agent *a*
Doing action A is necessary for agent *a* to carry out goal G
Therefore agent *a* ought to do action A.

and the *sufficient condition scheme* (W2):

W2 G is a goal for agent *a*
Doing action A is sufficient for agent *a* to carry out goal G
Therefore agent *a* ought to do action A.

Walton associates four critical questions with each of these schemes:

CQ1 Are there alternative ways of realising goal G?
CQ2 Is it possible to do action A?
CQ3 Does agent *a* have goals other than G which should be taken into account?
CQ4 Are there other consequences of doing action A which should be taken into account?

The treatment of practical reasoning as argument schemes and critical questions lends itself nicely to the defeasible nature of reasoning about action. In addition to his schemes for practical reasoning Walton also details 25 other argument schemes, which are presented as a classification in [164]. Other such typologies of argument schemes of varying sizes have also been given by Kienpointner [93], Grennan [76], and Katzav and Reed [91], among others. However, as the subject dealt with in this thesis solely concerns the topic of practical reasoning I shall only refer here to Walton's schemes for practical reasoning. In Chapter 3 I will examine in more detail the second of these

[4] In this and the next schema, I label each of Walton's symbols for clarity. In an earlier account [163], Walton gave a more detailed version of these schemes. The earlier scheme has five rather than four critical questions associated with it, with the fifth question enquiring whether or not the action is acceptable/"the best alternative".

schemes for practical reasoning, W2. There I will highlight some issues associated with this scheme and also propose an extension to it which will form the underlying basis for my theory of persuasion in practical reasoning.

Walton views argument schemes as ways of representing arguments embedded within dialogues. Together with Krabbe he has provided a typology of the different dialogues that can be used in human communication, which I discuss in the next subsection.

2.4.4 Walton and Krabbe's Dialogue Typology

In [167], Walton and Krabbe have identified a number of distinct dialogue types used in human communication: Persuasion, Negotiation, Inquiry, Information-Seeking, Deliberation, and Eristic Dialogues. This typology has proved to be influential in the study of argumentation theory and its application to agent systems (though Walton and Krabbe make no claims for its comprehensiveness). All these types are characterised by their initial positions, main goal and the aims of the participants. They are summarised below in Table 2.1, which is taken from [167].

Table 2.1: **Types of Dialogue**

Type	Initial Situation	Main Goal of Dialogue	Participants' Aims
Persuasion	Conflicting points of view	Resolution of such conflicts by verbal means	Persuade the other(s)
Negotiation	Conflict of interests and need for cooperation	Making a deal	Get the best out of it for oneself
Inquiry	General ignorance	Growth of knowledge and agreement	Find a proof or destroy one
Information-Seeking	Personal ignorance	Spreading knowledge and revealing positions	Gain, pass on, show or hide personal knowledge
Deliberation	Need for action	Reach a decision	Influence the outcome
Eristic Dialogue	Conflict and antagonism	Reaching an accommodation in a relationship	Strike the other party and win in the eyes of onlookers

The Walton and Krabbe descriptions are summarised as follows (in the order of [167]):

- A **Persuasion** dialogue involves an attempt by one participant to have another participant endorse some proposition or statement. The statement at issue may concern the beliefs of the participants or proposals for action[5], and the dialogue

[5] In Walton's [166], (which was published after the book with Krabbe [167]), he states that persuasion in his model involves arguments "to show or prove to the respondent that [a] thesis is true" [p. 37]. He also

may or may not involve conflict between the participants. If the participants are guided only by the force of argument, then whichever participant has the more convincing argument, taking into account the burden of proof, should be able to persuade the other to endorse the statement at issue.

- A **Negotiation** dialogue occurs when two or more parties attempt to jointly divide some resource (which may include the participants' own time or their respective capabilities to act), where the competing claims of the participants potentially cannot all be satisfied simultaneously. Here, co-operation is required by both parties in order to engage in the negotiation dialogue, but, at the same time, each participant is assumed to be seeking to achieve the best possible deal for him or herself.

- An **Inquiry** dialogue occurs when two or more participants, each being ignorant of the answer to some question, and each believing the others to be ignorant also, jointly seek to determine the answer. These dialogues do not start from a position of conflict, as no participant has taken a particular position on the question at issue; they are trying to find out some knowledge, and no one need resile from their existing beliefs. Aircraft disaster investigations may be seen as examples of Inquiry dialogues.

- An **Information-Seeking** dialogue occurs when one party does not know the answer to some question, and believes (perhaps erroneously) that another party does so. The first party seeks to elicit the answer from the second by means of the dialogue. Expert consultation is a common important subtype of this type of dialogue. Note, when the information sought concerns an action or course of action, I term this type of dialogue, a **Plan-Seeking** dialogue.

- A **Deliberation** dialogue occurs when two or more parties attempt to agree on an action, or a course of action, in some situation. The action may be performed by one or more the parties in the dialogue or by others not present. Here the participants share a responsibility to decide the action(s) to be undertaken in the circumstances, or, at least, they share a willingness to discuss whether they have such a shared responsibility.

- An **Eristic** dialogue is one where the participants vent perceived grievances, as in a quarrel, and the dialogue may act as a substitute for physical fighting.

Most human dialogues are in fact combinations of these ideal types. For example, a debate may contain persuasion, information-seeking and deliberation in turn, each

states that arguments about actions fall under the deliberation dialogue type which is characterised by "the need to take action to solve a problem or generally move ahead in some practical sphere" [p. 151]. It is also under deliberation dialogues where Walton discusses practical reasoning, presumptive arguments and his schemes for practical reasoning. Examples of the form "We should take action A" are given in [166] under the deliberation category e.g., "cigarette tax should not be reduced" [p. 159].

embedded in the larger interaction. Moreover a dialogue may shift between types as it proceeds. This dialogue typology has proved to be of importance in argumentation theory and its application to AI for a number of reasons, including:

1. By identifying a dialogue as falling under one of the particular types, the participants are aware of the goal they are trying to achieve by engaging in the dialogue interaction.

2. Shifts that occur during the course of an interaction may lead to fallacies (if the shift is illicit) and misunderstandings (if the shift goes unnoticed).

3. The pragmatic meaning of speech acts, e.g., ‘assert’ or ‘inform’, is determined by the dialogue type.

With the exception of eristic dialogues (which are generally viewed as being beyond rational discourse), I have taken the above dialogue types and given a more precise characterisation to them. This is done using the initial beliefs and aims of the participants and the ways in which these can change during the course of the dialogue. This allows us to identify any shifts in the dialogue type, and the changes which the parties can make to reach agreement.

Tables 2.2 – 2.4 show the analysis for three typical situations. Table 2.2 shows the possibilities where two parties discuss their beliefs regarding a single proposition. Table 2.3 shows the possibilities when two parties discuss whether a particular action should be performed or not. Table 2.4 shows the situation where two parties discuss the performance of either, both or neither of two actions, which may be performed.

Table 2.2: **Model of a Discussion Over Beliefs**

A/B	B believes *p*	B believes ¬ *p*	B believes *p* or ¬ *p*
A believes *p*	Agreement	Disagreement or Persuasion	B info seeks A
A believes ¬ *p*	Disagreement or Persuasion	Agreement	B info seeks A
A believes *p* or ¬ *p*	A info seeks B	A info seeks B	Inquiry

Table 2.3: **Model of a Discussion Over Actions**

A/B	Do *p*	Do not do *p*	Do *p* or do not do *p*
Do *p*	Agreement	Disagreement or Persuasion	B plan seeks A
Do not do *p*	Disagreement or Persuasion	Agreement	B plan seeks A
Do *p* or do not do *p*	A plan seeks B	A plan seeks B	Deliberation

Table 2.4: **Model of a Discussion Over Multiple Actions**

A wants/B wants	A does *p* and *q*	A does *p* but not *q*	A does *q* but not *p*	A does not do *p* or *q*	B has no opinion
A does *p* and *q*	Agreement	Conflict or Persuasion	Conflict or Persuasion	Conflict or Negotiation	Plan seeking
A does *p* but not *q*	Conflict or Persuasion	Agreement	Conflict or Persuasion	Conflict or Persuasion	Plan seeking
A does *q* but not *p*	Conflict or Persuasion	Conflict or Persuasion	Agreement	Conflict or Persuasion	Plan seeking
A does not do *p* or *q*	Conflict or Negotiation	Conflict or Persuasion	Conflict or Persuasion	Agreement	Plan seeking
A has no opinion	Plan seeking	Plan seeking	Plan seeking	Plan seeking	Deliberation

These tables model the space of all possible dialogue types appropriate to these situations. For example, consider Table 2.2 and the situation where A believes *p* and B believes $\neg p$, intersecting in the second row and third column of the table. Here we have a disagreement between the two parties as to the truth of proposition *p*. If A can successfully persuade B that *p* is the case, then the situation is represented as one shift left in the columns where agreement is reached that *p* is the case. However, if the opposite were to occur where B successfully persuades A that $\neg p$ is the case, then one shift down the rows would occur where agreement is reached that $\neg p$ is the case. Now consider the case where party A is committed to $\neg p$ being the case, but party B is unsure as to whether *p* or $\neg p$ is the case. This situation is represented in the cell where the third row and last column intersect. In this case B must seek information from A, who will respond that $\neg p$ is the case. If B chooses to believe this then a shift in columns to the left will occur, resulting in the situation where agreement is reached. Finally, in the situation where neither A nor B are sure as to whether *p* or $\neg p$ is the case we have in inquiry dialogue, as shown at the intersection of the last row and last column of the table. If and when one party commits to believing *p* or $\neg p$ then a shift will occur in the appropriate direction and whichever scenario results will then apply, as described above.

Table 2.3 models the same scenarios though the difference here is that the two parties are discussing whether a particular action should be performed. This means that the dialogues involved in this table are slightly different from those involved in the previous table. In Table 2.3 when there is a situation where one party is committed to one of the actions being performed and the other party commits to neither action then we have a plan-seeking dialogue. In the situation where neither party is committed to an action then we have a deliberation dialogue.

The situation becomes a little more complex when we consider discussions over multiple actions, as depicted in Table 2.4. In this table the discussion between the two parties A and B is concerned with the number of actions that party A should do. Agreement is reached in such a discussion when both parties agree upon the actions that A should do. For example, where both parties think that A should do both actions *p* and *q*, as depicted in the cell intersecting the second row and second column, we have agreement. Likewise, we have agreement when both parties agree that A should do *p* but not *q*, as depicted in the cell intersecting the third row and third column. Conflict occurs when division of the actions is not agreed upon. However, there are two different kinds of conflict here. In the first case, both parties agree upon the allocation of one action but not the second action. For example, as shown in the cell intersecting the second row and third column where both parties agree that A should do *p* but they have opposing views as to whether A should also do *q*. To reach agreement in this case, one party must persuade the other to change their view so the division of the actions is agreed upon. For example, if party B can be persuaded to change his mind and accept that A does *p* and *q* then a shift left will occur and agreement will be reached. In the second case where there is conflict, both parties have directly opposing opinions as to the division of the actions. For example, as shown in the cell intersecting the second row and fifth column where A wants to do *p* and *q* but B wants A to do neither *p* nor *q*. Where there is this type of conflict it is not a persuasion dialogue that is needed but a negotiation one, as there is currently no agreement on the division of the actions and a trade-off is possible. We may also have a situation where one party has no opinion as to the division of the actions and the other party does, then we have a plan- seeking dialogue. For example, when B has no opinion as to the division of the actions but A wants to do *p* and *q*, we are in the cell where the second row intersects with the sixth column. Finally, in the situation where neither party is committed to an action then we have a deliberation dialogue.

Representing the dialogues in this way leads to a number of observations relating to reaching agreement:

- we can see the space of possible moves available to the participants;
- we can see how agreement can be reached;
- we can see how many changes are needed if agreement is to be reached;
- we can see which participant must change if agreement is to be reached.

Providing the information in the form of these matrices gives a more structured and precise characterisation of the dialogue types than the informal descriptions of [167]. When the participants have a clear understanding of the gaps between their positions the task of deciding what shifts in position they should try to induce, or may need to make, is facilitated. Of course, whether a party is willing to change his position will depend on his other beliefs, and the utility he ascribes to actions and the states resulting from action. The structures, however, do provide a basis for forming strategies and heuristics to inform the conduct of the various types of dialogue. These matrices are intended as a structure by which we can model the dialogues from Walton and Krabbe's typology and shifts that can occur between the dialogue types. However, the work presented in the forthcoming chapters of this thesis focuses on one dialogue type in particular: persuasion dialogues regarding actions. Additionally, there is some recent work by Cogan *et al.* [46] which also aims at refining the definitions of the dialogue types in the Walton and Krabbe typology.

This concludes my discussion of argumentation theory in philosophy and I will now examine how this field has been applied to AI in recent years.

2.5 Argumentation in AI

Over the last decade theories from the field of argumentation, as described in the previous section, have been applied to the design of reasoning mechanisms for multi-agent systems. In this section I give a general overview of this application and discuss some important concepts from this field that are applicable to the work presented in this thesis.

2.5.1 Overview

It has only been within recent years that computer scientists have taken the large body of work from argumentation theory and applied it to the area of multi-agent systems. With the ongoing development of multi-agent systems it is natural that agent designers have the need to equip autonomous agents with mechanisms by which they can reason and argue with themselves and other agents. As discussed in Section 2.4, argumentation provides a means by which uncertain and incomplete information can be reasoned about. This is of obvious advantage to agent systems as it is typically the case that agents will be operating in domains with incomplete and uncertain knowledge and they will often need to make decisions whilst being aware of this fact. In such a situation argumentation can play an important role by providing tentative conclusions for or against a claim in the absence of further information to the contrary of the claim.

The models of argumentation for AI need to include some method by which an agent can assess the relative worth of the arguments pertinent to a particular debate, i.e., the agent needs to be able to determine which arguments are the most convincing. One such method has been proposed by Dung [55] in which an argument for a claim is accepted or rejected on the basis of how well it defends itself against other arguments that can attack and defeat it. Dung's system is an example of an Argumentation Framework and it has proved to be a particularly influential formal system of defeasible argumentation. This system, and a particular extension to it, is of importance in the work presented in Chapter 5 onwards of this thesis and I shall discuss Dung's system in more detail in Section 2.5.3 and Section 5.5. Before this I shall comment on a few other recent and notable developments in the application of argumentation to AI.

The incorporation of argumentation into AI is done through formal representation of the features of argumentation, which themselves derive from the study of informal logic. The application of argumentation to multi-agent systems was first introduced in 1998 in a paper by Parsons *et al.* [120]. Since then, there has been increasing interest in the application of argumentation to a number of sub-fields of multi-agent systems. One such sub-field from the multi-agent systems community that has benefited from argumentation theory is that which deals with the design of negotiation models for agents [60, 94, 96]. The need for negotiation arises from the fact that agents are often dependant upon one another to fully complete all their tasks. Such social ability is one of the capabilities cited in [169] that is deemed necessary for an agent to have in order to be classified as 'intelligent'. Recall from Section 1.2, the definition given in [169] for social ability in autonomous agents is: "Intelligent agents are capable of interacting with other agents (and possibly humans) in order to satisfy their design objectives." However, during any interaction it is not always the case that agents' knowledge, interests, preferences and goals will all be overlapping and so situations of conflict can easily arise. In order to try and resolve any conflicts that arise, agents need to be equipped with mechanisms to aid conflict resolution. The way in which conflicts are often resolved in everyday life is through the exchange of reasoned argument and justification of a stance. So, agent designers have looked to argumentation theory for help in this matter. One particular approach that has been proposed in an attempt to aid resolution of conflicts through argumentation and to enable agents to negotiate more efficiently has been professed by Rahwan *et al.* [134]. Their method is known as *argumentation-based negotiation* and the authors claim that the approach allows for more sophisticated forms of negotiation to take place than have been previously proposed. Rahwan *et al.* provide the background motivation for the need for such an approach in [134] by discussing the shortcomings of existing approaches to negotiation, including game-theoretic and heuristic-based approaches. They also give full details of the elements that comprise argumentation-based negotiation frameworks. The intuition behind argumentation-based negotiation is that agents may increase the likelihood and quality of an agreement by exchanging arguments which influence each others' states. Additionally, it is stipulated that the exchange of arguments is sometimes essential to this process when various assumptions about agents' rationality do not suffice. In making use of argumentation within a negotiation setting the designers of these models are aiding the conflict resolution process by enabling participating agents to exchange arguments in an attempt to persuade their counterparts to cooperate more with them. Thus, the concept of persuasion also plays an important role in such exchanges [137]. However, the notion of persuasion as expressed in the work in this thesis is not intended for use in a negotiation scenario. That is not to say that there are not overlapping features of the two models, but the concept of persuasion presented here is intended solely as a basis to enable persuasive argument about action to take place, outside of a negotiation scenario. In particular, trade-offs and bargaining are not considered.

In the same strand of work, another model of negotiation that makes use of argumentation is the interest-based negotiation model of Rahwan [133]. Interest-based negotiation involves argumentation relating to agents' underlying interests, in a negotiation setting. It enables agents to exchange arguments as to why a goal is desirable, in addition to exchanging arguments about beliefs. The idea is that if an agent is aware of the reasons as to why his opponent has adopted a particular goal, then he will be able to engage in a discussion with his opponent regarding the suitability of his goals. Thus, Rahwan's model for interest-based negotiation includes elements to enable the exchange of arguments about goals to take place, within a negotiation setting. The ability to use argument enables agents to discuss the underlying motivations and interests for the adoption of their goals, in order to try and exert influence upon each other's preferences [135]. The exchange of arguments about goals is an aspect of multi-agent argumentation that is also catered for by the theory presented in this thesis, as will become clear later on. Additionally, in the theory I present arguments are also permitted, and indeed encouraged, that relate to components of the practical reasoning model other than beliefs. Rahwan's method of interest-based argumentation is clearly stated in [133] as intended solely for agents interacting within a negotiation scenario in order to reach agreement over the exchange of scarce resources. Thus, it does not serve the same purpose as the theory I propose.

A number of systems have been developed to implement a negotiation process within an argumentation setting. One of the earliest examples of such a system is Sycara's PERSUADER system [153], which is also described in [169]. This system provides a framework for conflict resolution through agent supported negotiation and mediation. The particular domain that PERSUADER is set in is labour management dispute, where there are three parties involved in the negotiation: a labour union, a company and a mediator. The PERSUADER system acts as a mediator through the exchange of proposals and counter-proposals to facilitate the disputants' problem solving so that a mutually-agreed-upon settlement can be achieved. The system is based upon a negotiation model that can handle multiple issues and it can order and present arguments for a particular position according to their strength on a pre-defined scale. PERSUADER is a mediation system designed to enable one agent to reason about the different proposals presented to it. In this sense it does not constitute a full multi-agent system, though it does nonetheless make use of a form of negotiation based upon the exchange of arguments.

All the examples listed above make use of argumentation to facilitate resolution of conflict in a negotiation situation. Due the theory presented in this thesis being within the sphere of persuasion in practical reasoning, I shall not consider the above methods for argumentation about negotiations any further.

2.5.2 Argumentation for eDemocracy

I now focus attention on another particular domain that lends itself well to the application of AI and has only emerged in relatively recent years: eGovernment and eDemocracy. This domain has emerged due to the ubiquity of computer systems in the public domain and the interconnectivity afforded through inexpensive public access to the Internet. eGovernment initiatives comprise a number of objectives that aim to automate government systems and enable members of the public to interact with the Government through the provision of electronic media. AI and argumentation can assist in such objectives. A relatively early system to facilitate online discussion between citizens and representatives of public interest groups was the Zeno Argumentation Framework of Gordon and Karacapilidis [71]. Zeno embodies a formal model of argument that is based upon the informal model of Rittel called the Issue-Based Information System (IBIS) [144]. Zeno is intended as a mediation system of an electronic discussion forum which incorporates support through argumentation. Its main feature is a type of labelling function to represent arguments so that the relationship of positions regarding a solution to a practical issue can be assessed. In this way, a dialectical graph can be constructed showing the pros and cons of the choices available, in order to decide upon a solution to a practical issue. Users are able to express their preferences for particular choices and provide qualifications for these preferences and thus the quality of opposing options can be assessed. Zeno does indeed provide an effective method of analysing the pros and cons of arguments for a particular choice on a particular issue. However, one obstacle presented by the system is that the framework was deemed too difficult to be effectively used by laypersons. However, since its introduction in [71], Zeno has since been further developed and has been used in several eDemocracy pilot applications [72]. Zeno is just one example of a computer system designed to mediate online dialogues between users. A number of other systems with similar aims have recently been developed to meet the objectives of eGovernment, with the DEMOS system of Lührs *et al.* [101] being one such example. However, I conclude the discussion of the application of argumentation in eGovernment at this point as I will return to it in Chapter 6 where I discuss my own mediation system for eDemocracy based upon my model of practical reasoning, along with a deeper background discussion of the issues related to the subject.

In the next subsection I examine one of the more abstract and general methods established to evaluate arguments.

2.5.3 Argumentation Frameworks

As acknowledged in Section 2.5.1, there has been considerable interest in the field of argumentation in the representation of arguments through abstraction mechanisms named 'argumentation frameworks'. One influential contributor to this area is Dung who introduced his formal system for evaluating arguments in [54, 55] and this system has provided the basis for much of the subsequent work in this area. In [55] Dung defines an Argumentation Framework as a finite set of arguments X, and a binary relation between pairs of these arguments called *an attack*. These relationships are modelled as directed graphs showing which arguments attack one another. No concern is given to the internal structure of the arguments, so the status of an argument can be evaluated by considering whether or not it is able to defend itself from attack from other arguments with respect to a set of arguments $S \subseteq X$. Following this definition, a given argument A is said to be 'acceptable', with respect to the set S, if any attack, by an argument $B \in X$, on A is itself attacked by some other argument C which is in the set S, i.e., acceptable arguments in S are those which are arguments that are defended by arguments within S against all attacks from other members of the set X. Additionally, the set S is said to be 'conflict-free' if there is no pair of arguments A, B in S such that B attacks A. Furthermore, a set of arguments S is defined as being 'admissible' if it is conflict-free and each argument it contains is acceptable with respect to S, i.e all arguments in the set are defended against all attacks, from arguments in X but not in S. From these definitions Dung goes on to define the semantics of such argumentation frameworks through the notion of a *preferred extension*. The preferred extension of an argumentation framework is defined as the maximal admissible set of arguments in the particular framework. This preferred extension allows us to determine the maximally consistent set of beliefs, with respect to the arguments in the given set. Other additional semantics defined by Dung that can follow from this are *grounded semantics*, where arguments are not permitted to defend themselves against attacks from outside the set, and *stable semantics*, where every argument that is not in the preferred extension is attacked by some argument in the preferred extension. The formal description for each of these definitions can be found in [55], and due to argumentation frameworks being used in the work presented in later chapters of this thesis, an outline of these formal definitions is also provided in Appendix C.

A number of extensions and variations to Dung's model have been proposed with [2, 35, 95, 162] being notable examples. Here I shall focus on one particular example, namely the Value-Based Argumentation Frameworks (VAFs) of Bench-Capon [26], due to their applicability to the work on values in this thesis. VAFs are essentially an extension to Dung's argumentation frameworks to allow arguments to be evaluated according to the values that they promote. By the inclusion of values in the analysis of the acceptability of arguments VAFs enable distinctions to be made between different audience's preferences, in the sense of Perelman's description given in Section 2.1.1. Whereas in Dung's frameworks an argument is always defeated by an attacker, unless that attacker can itself be defeated, in VAFs, attack is distinguished from *defeat for an audience*. This allows a particular audience to choose to reject an attack, even if the attacking argument cannot itself be defeated, provided that audience ranks the purpose motivating the attacked argument – the value cited in its justification – as more important than that motivating the attacker. Values also relate arguments, so that decisions made about one attack will decide other attacks in the framework. Within a VAF, therefore, which arguments are accepted depends on the ranking that the audience (characterised by a particular preference ordering on the values) to which they are addressed gives to these motivating purposes. However, one point to note is that if the attacker promotes the same value as the argument which it is attacking, the attack always succeeds.

The preferred extension of a VAF for an audience is the maximal subset S of $Args$ such that no argument in S defeats any other argument in S given the value ordering of that audience, and all arguments in S are acceptable to that audience with respect to S, i.e., for any argument A in S, if A is defeated by an argument A' that is not in S, then there exists an argument in S that defeats A' on the given value ordering. Given the definitions for VAFs we can also determine the preferred extension of a VAF with respect to an audience. Thus, the preferred extension represents the maximal consistent set of acceptable arguments with respect to the argumentation framework and a given value ordering, which is the maximal consistent position for an audience with that value ordering. In [26] Bench-Capon further shows that the preferred extension for a given value ordering is unique and non-empty, provided it contains no cycles in which every argument relates to the same value. Conversely, in Dung's framework there may be multiple preferred extensions, which lead to algorithms for determining the preferred extension to be intractable, as shown in [56]. A Value-Based Argumentation Framework with a given ordering on values can be mapped to a pair $(Args, Defeat)$, where '*Args*' is the set of arguments in the framework and '*Defeat*' is the subset of attacks which succeed for that audience. If there are no cycles in a single value in the VAF, the resulting argumentation framework is cycle-free, so we can determine which arguments in $Args$ are acceptable to a particular audience by determining the preferred extension for that audience using polynomial time algorithms [26]. So, VAFs provide a mechanism by which arguments can be evaluated against each other, whilst also accounting for the fact that different audiences may find opposing arguments acceptable based upon their individual value preferences. As indicated earlier in Section 2.1.1, the notion of value plays an important and meaningful role in the ideas presented in this thesis and this will become clearer in the ensuing chapters. Chapters 5, 6, 7 and 8 further discuss the use of VAFs and show how they can be used in accordance with my model of persuasion over action to allow a BDI agent to determine the best course of action to taken, given its value preferences. The full description and formal definitions for Value-Based Argumentation Frameworks can be found in [26], and due to their use in Chapters 5, 6, 7 and 8, I also provide a summary of the formal definitions of VAFs in Appendix C.

This concludes my discussion of argumentation in AI, which has highlighted particular aspects of the field that are applicable to the ideas presented in this thesis. More detailed accounts of argumentation in AI can be found in a survey given by Carbogim *et al.* [42] and in a more recent survey conducted by the Argumentation Service Platform with Integrated Components Project (ASPIC) [4].

I will now examine how arguments can be exchanged through dialogue by surveying current methods used for agent communication and interaction.

2.6 Agent Communication and Interaction

In this section I examine the field of agent communication. I discuss the main methods and formalisms that have been developed to enable autonomous agents to engage in structured dialogue and exchange arguments. I discuss the problems associated with the main proposals and how these issues are being addressed in current research.

2.6.1 Dialogue Games

Dialogue games have been studied extensively in the philosophical literature since the time of Aristotle [6] and also more recently in philosophy [80], computer science and AI [24, 116]. During a dialogue game the participating players exchange utterances, known as ‘moves’ or ‘locutions’, according to a set of defined rules known as a ‘dialogue game protocol’. Each move has an identifying name associated with it and its contents are some statement (represented in a suitable langauge) which contribute to the dialogue. Such moves are exchanged by participants until the dialogue terminates, according to some termination rules.

One particular element that features in a number of dialogue games is that of a *commitment store*. This notion derives from Hamblin’s study of fallacious reasoning in [80]. A commitment store can be viewed as a repository that records the statements made by participants which incur commitments. This makes a distinction between the notions of belief and commitment, whereby commitment in a dialogue may not incur commitment outside the dialogue, or reflect the participants’ true beliefs. This is of importance in analysing dialogues since they are based upon verifiable records of statements made, rather than participants’ internal beliefs. However, it is important to note that there are a number of different notions of ‘commitment’ in the literature. Firstly, there is Hamblin’s notion [80] whereby incurring commitment in the context of a dialogue does not automatically incur commitment outside the dialogue. To quote Hamblin:

> "The [commitment] store represents a kind of *persona* of beliefs: it may not correspond with his [the speaker's] real beliefs, but it will operate in general, approximately as if it did." [80, p. 257].

Secondly, there is Walton and Krabbe's notion [167] whereby commitment incurred in a dialogue may not only relate to statements within the dialogue but it may also incur commitment to action outside of the dialogue. To quote:

> "... we came to think of commitment as a practical idea, one that has to do with imperatives directing an agent to a course of action in a particular situation." [167, p. 7].

They further go on to state that:

> "Propositional commitments, such as those of which Hamblin's commitments stores were supposed to keep track, constitute just a special case of commitment to a course of action..." [167, p. 8].

Finally, in the multi-agent systems sense a commitment is often regarded as a persistent goal that the agent is trying to achieve [47], and this may be extended to cover the notion of social commitment whereby an agent is committed to doing something for another agent or committed to partaking in a particular social role [43].

In general, commitment stores in dialogue games and protocols are 'publicly' accessible (as is the case for the protocol proposed in this thesis in Chapter 4) and they can be viewed by all participants of the dialogue, thus aiding the parties, and any referee, in identifying inconsistencies between participants' beliefs. Contrary to publicly visible commitments, Walton and Krabbe also discuss in [167] what they term as *dark-side commitments*. These are commitments that are in general not known to the players, neither the holder of the commitment nor the other participants of the game, but they may be revealed during the course of the dialogue game. The use of commitment stores in dialogue games has proved useful in the specification and verification of semantics for such games, as I will discuss in Section 2.6.2. Importantly, these semantics need make no appeal to the internal (and hence possibly inaccurate) states of agents.

Dialogue games have been used in philosophy to demonstrate how fallacious arguments can be identified. A large body of research has been conducted into fallacious arguments dating back to Aristotle's work *On Sophistical Refutations*, as discussed by Hitchcock in [81], to more recent work on the topic such as the pioneering study of logical fallacies conducted by Hamblin and presented in [80]. Here, to serve as an example of a typical dialogue game, I shall examine one particularly well-known dialogue game: MacKenzie's DC [102]. This dialogue game was constructed to address a particular form of fallacious argument – 'begging the question' – and it has been discussed by numerous other people such as Moore [116] and Bench-Capon *et al.* [28], among others. The game addresses this particular fallacy by defining rules to enable a player to be committed to a question, in addition to being committed to a statement. The motivation for MacKenzie's game was to develop a set of rules to enable a dialogue about some argument to be conducted between two participants and the rules would guarantee that a resolution to the argument would be reached. The game commences with one of the participants asserting a claim and thus incurring a commitment to this proposition. This proposition may be questioned, challenged and defended, through each participant taking turns to make one of the moves of the game. The rules of the game ensure that circularity in the arguments is avoided through the use of a *resolution demand* rule, where either the proponent must withdraw the proposition or the opponent accept it, in the light of any apparent contradiction that has arisen. A summary of the DC system [102] is given below:

DC consists of five possible move types as follows:

1. Statements, i.e., 'P', 'Q' etc., and truth functional compounds of statements, such as 'Not P', 'If P then Q', etc.

2. Withdrawals, i.e., withdrawal of the statement P is 'no commitment to P'.

3. Questions, i.e., 'Is it the case that P?'.

4. Challenges, i.e., 'Why P?'.

5. Resolution demands, i.e., Resolution demand of P is 'Resolve whether P'.

In addition to the above move types, there are also six rules to regulate commitment stores:

1. Statements: After a statement 'P', unless the preceding event was a challenge, 'P' is included in both participants' commitments.

2. Defences: After a statement 'P', when the preceding event was 'Why Q?', both 'P' and 'If P then Q' are included in both participants' commitments.

3. Withdrawals: After the withdrawal of 'P', the statement 'P' is not included in the speaker's commitment. The hearer's commitment is unchanged.

4. Challenges: After the challenge of 'P', the statement 'P' is included in the hearer's commitment; the statement 'P' is not included in the speaker's commitment; and the challenge 'Why P?' is included in the speaker's commitment.

5. Questions and resolution demands: These locutions do not themselves affect commitment.

6. Initial Commitment: The initial commitment of each participant is null.

In addition to the above rules regarding commitment stores there are also eight further ‘dialogue rules’:

1. Each participant contributes a locution at a time, in turn; and each locution must be either a statement, or the withdrawal, question, challenge or resolution demand of a statement.

2. No statement may occur if it is a commitment of both speaker and hearer at that stage.

3. A conditional whose consequent is an immediate consequence of its antecedent must not be withdrawn.

4. After ‘Is it the case that P?’, the next utterance must be either ‘P’, ‘Not P’ or ‘No commitment P’.

5. A conditional whose consequent is an immediate consequence of its antecedent must not be challenged.

6. After ‘Why P?, the next utterance must be either; (i) ‘No commitment P’; or (ii) The resolution demand of an immediate consequence conditional whose consequent is ‘P’ and whose antecedent is a conjunction of statements to which the challenger is committed; or (iii) A statement not under challenge with respect to its speaker (i.e., a statement to whose challenge its hearer is not committed).

7. The resolution demand of ‘P’ can occur only if either; (i) ‘P’ is a conjunction of statements which are immediately inconsistent and to all of which its hearer is committed; or (ii) ‘P’ is of the form ‘If Q then R’, and ‘Q’ is a conjunction of statements to all of which its hearer is committed; and ‘R’ is an immediate consequence of ‘Q’; and the previous event was either ‘No commitment R’ or ‘Why R?’.

8. After ‘Resolve whether P’, the next utterance must be either; (i) The withdrawal of one of the conjuncts of ‘P’; or (ii) The withdrawal of one of the conjuncts of the antecedent of ‘P’; or (iii) The consequent of ‘P’.

One of the main features of MacKenzie’s DC is that it enables errors within logical reasoning to be detected and it has received considerable attention from both the philosophy and AI communities. In [116] Moore gives a more detailed description and discussion of the DC game, as well as a computational model for the game. Moore's version has been subsequently developed and refined by Yuan in [172].

Most dialogue games are based upon a similar structure to DC, comprising players, locutions/moves made up of syntax and semantics, rules of interaction and some form of commitment store (or similar structure) to record commitments made by the players. A dialogue game protocol for the theory of practical reasoning presented in this thesis is given in Chapter 4.

In addition to DC, many other dialogue games have been proposed that are based upon more expressive argument structures. One prominent example is the "Toulmin Dialogue Game" [24], based upon Toulmin's argument schema [157], which was discussed in Section 2.4.1 of this chapter. An implementation of the Toulmin schema in the form of a dialogue game has been undertaken in [24] which thus provides an example of a dialogue game based on an argument scheme (as is the dialogue game developed later in this thesis). The application used in that particular example is to facilitate the conduct of effective legal reasoning. A number of proposals have been given for systems to model legal reasoning in the form of dialogue games. A seminal piece of work in this area is Gordon's "*The Pleadings Game*" [70]. The motivation for this particular example was an attempt to model the process of legal pleadings, whereby pleadings entail a pre-trial process in which the disputants identify points of agreement and the issues to be resolved at trial. In this game Gordon characterises the process of legal pleadings as a two player dialogue game, designed to identify which issues were agreed between the parties, and which remained in dispute and so required decision in a trial. A further example of a dialogue game for legal applications is Lodder's DiaLaw [100]. His game is based upon Hage's 'Reason Based Logic' (RBL), a non-monotonic logic that was developed to deal with legal rules [79].

Dialogue games have proved to be a useful way in which we can model and reason about exchanges of information between participants, in a number of different domains, including law, philosophy and AI. A comprehensive discussion of Dialogue Game Theory in general, plus a number of examples of dialogue games and systems can be found in [116]. I return to the area of dialogue games in Chapter 4 of this thesis where I propose a dialogue game protocol to model how arguments can be exchanged as part of the practical reasoning process, and as embodied by the theory I present in Chapter 3.

I now examine some more specific approaches that enable dialogical exchanges to take place between autonomous agents.

2.6.2 Agent Communication

One major concern within the field of multi-agent systems is the need to develop efficient and expressive methods to allow intelligent agents to communicate with one another. This area has again benefitted greatly from established work in the discipline of philosophy, and in particular *speech act theory*. As discussed by Wooldridge in [169], speech act theory has its roots in the work of Austin [21] where communication is treated as a form of action which alters the physical state of the world and/or the mental state of the participants involved in the communication. Utterances made by participants are classified as *speech acts* which, according to Austin, have three different aspects as follows:

- the *locutionary act* which is the act of actually making an utterance (e.g., saying "Please pass me the salt"),
- the *illocutionary act* which denotes the action performed in saying something (e.g., "She requested that I pass the salt"),
- the *perlocution* which is the actual effect of the act (e.g., "She got me to pass the salt").

As Wooldridge explains [169], such speech acts are made through the statement of what Austin calls *performative verbs*, which correspond to various types of speech act, e.g., *inform, request, promise* etc. In addition to this classification, Austin also identified conditions that were required to be met in order for the execution of performative verbs to be successfully completed. He called these *felicity conditions* [21] and they are as follows:

- there must exist an acceptable conventional procedure for the performative, and the circumstances and persons must be as specified in the procedure,
- the particular persons and circumstances must be appropriate for the invocation of the procedure,
- the procedure must be executed correctly and completely by all participants. [21].

Austin's work on speech act theory was taken forward by Searle who in [145] identifies properties that must hold in order for a speech act between two participants (a speaker and a hearer) engaged in a communicative exchange to succeed. The classification of linguistic exchanges in speech act theory has proved to be extremely influential upon the development of communication mechanisms for autonomous agents. One of the first major applications of speech act theory to agency was produced by Cohen and Perrault [49] in the late 1970s. They give an account of the semantics of speech acts whereby the properties of an action are defined in terms of pre-conditions and post-conditions. This method enables speech acts to be reasoned about and deployed in communicative exchanges between agents. This work was further developed by Cohen and Levesque who specified a theory of intention which allowed speech acts to be modelled as actions performed by agents in order to fulfil their intentions [48]. Further enhancements to this theory have since been made by Bretier and Sadek [38] who provided an implementation of the theory to be used in rational agents. Before I discuss particular examples of agent communication languages (ACLs) I will first examine the different semantics that can be used to define the meaning of an ACL. As in human communication, languages for autonomous agents comprise a syntax, which defines the legal constructs of the language, and semantics, which define the meaning of the language. A number of different types of semantics have been identified by which agent communication languages can be defined and the protocol detailed in Chapter 4 of this thesis makes use of some of these types of semantics. Thus, I provide a brief summary of three different kinds of semantics by which programming languages for agent communication can be defined, as given in van Eijk's PhD thesis on the topic [58].

The first type is an axiomatic semantics in which the meaning of the speech acts are not explicitly specified, but are given in terms of the properties that the language concepts satisfy. These semantics are generally defined by three elements: pre-conditions, which are constraints that must be met in order for the speech act to be made, post-conditions, which are brought about by the speech act being made, and the speech act itself. Later on in this section I will discuss two particular ACLs, FIPA ACL [61] and KQML [121], which are both defined in terms of axiomatic semantics. One manner by which axiomatic semantics for dialogue game protocols can be specified is through the use of pre- and post-conditions with respect to players' *commitment stores*, which were discussed in Section 2.6.1. Here pre-conditions for a move may be verified against what is currently held in the players' commitment stores and post-conditions enforce updates on commitments to be added to the store. The use of such commitment stores thereby provides a *public* semantics. Alternative axiomatic semantics can also been given using agents' internal states, giving rise to a *private* semantics. Axiomatic semantics provide sufficient specification to enable implementation of an agent communication language/protocol. However, according to van Eijk [58], in order to give a more precise meaning to an ACL and to enable us to study its formal properties, we can supplement an axiomatic semantics with another form of semantics, which I will now describe.

The second approach to ACL semantics discussed by van Eijk is an operational one. This type of semantics views moves in a dialogue game as sequences of transitions operating upon some abstract state machine. An example of an operational semantics for an ACL is provided by van Eijk *et al.* in [59]. The advantage of using an operational semantics is that implementation of the language is facilitated, as it can be based upon an implementation of the corresponding abstract state machine.

The third and final type of semantics discussed by van Eijk is a denotational semantics. This type defines the moves of the dialogue by assigning an abstract mathematical meaning (called a denotation) to each one. In this way the meaning of each individual move can be defined and studied in isolation. An example of how denotational semantics can be used in the definition of dialogue game protocols is given in McBurney and Parson's semantics for deliberation dialogues [108].

This brief summary of semantics for use in ACLs is intended as the background motivation and justification for the semantics that I present as part of a persuasion protocol set out in Chapter 4. There I will give an axiomatic semantics for the protocol, and I also provide the outline of a denotational semantics for it in Appendix A. A more detailed discussion of the semantics of programming languages for agent communication and programming languages in general can be found in [58, 77, 156].

I now examine the two most prominent proposals for ACLs to date. These are the "Knowledge Query and Manipulation Language" (KQML) [121] and the "Foundation for Intelligent Physical Agents Agent Communication Langauge" (FIPA ACL) [61] and they have been greatly influenced by much of the work on speech act theory, as discussed earlier on in this subsection. I will now briefly describe each of these ACLs.

KQML was developed on the "Knowledge Sharing Effort (KSE)" project funded by the US Government's Defense Advanced Research Projects Agency (DARPA). The aim of the project was to develop protocols that would enable autonomous agents to exchange knowledge that was represented in a suitable format. The output of this effort was two connected languages: KQML and the Knowledge Interchange Format (KIF). KQML is defined as the message-based outer language of the communication, classifying messages into particular groups, called 'performatives', to establish a common format for the exchanges. Conversely, KIF is concerned only with providing a representation for the inner content of the communication i.e., the knowledge applicable in the particular domain. KQML proved to be influential amongst agent developers and it forms the basis of a number of implementations. However, numerous criticisms were directed at it on a number of grounds including its interoperability, lack of semantics and the omission of certain classes of messages to enable the expression of commitments. A more detailed description of the specifics of KQML and KIF, and a discussion of their criticisms, has been given by Wooldridge in [169].

The criticisms of KQML led to the development of a separate, though similar, agent communication language: FIPA ACL. In 1995 the Foundation for Intelligent Physical Agents developed their own ACL with the aim of establishing a standard communication language for use by autonomous agents. FIPA ACL is similar to KQML in that the syntax of both languages is similar, and also in that FIPA ACL, like KQML, uses an outer language to enable message passing of the separate inner content, which may be expressed in any suitable logical language. For the outer language, FIPA ACL provides 22 performatives to distinguish between the different kinds of messages that can be passed between agents. Three examples of FIPA ACL performatives are: *inform*, to pass information from one agent to another, *propose*, to offer a proposal for a particular executable action, and *request*, to ask for a particular action to be performed. The full specification of FIPA ACL performatives can be found in [61].

In order to avoid some of the criticisms that were directed against KQML the developers of FIPA ACL provided a comprehensive formal semantics for their language. These semantics made use of the work on speech act theory, as described earlier on in this subsection, through the definition of a formal language called *Semantic Language (SL)*. *SL* enables the representation of agents' beliefs, uncertain beliefs, desires, intentions and actions available for performance. To ensure that agents using the language are conforming to it, *SL* contains constraints (pre-conditions) in terms of formulae mapped to each ACL message that must be satisfied in order for compliance to hold e.g., agents must be sincere, and they must themselves believe the information they pass on to others. Additionally, *SL* enables the rational effects of actions (post-conditions) to be modelled, which state the intended effect of sending the message e.g., that one agent wishes another to believe some information passed from the first to the second.

Despite the enhancements that the FIPA ACL provided over KQML it has not escaped criticism itself. The main points of contention, as discussed by McBurney *et al.* [109], are summarised as follows:

- FIPA ACL was intended for purchase negotiation dialogues and has no means by which participants can state the purpose of alternative types of dialogue, e.g., inquiry, persuasion, etc.

- Its argumentative capabilities are severely limited with an under-provision of locutions to question and contest information. In addition to this, the agents' knowledge bases are assumed to be private, making resolution of conflicts difficult.

- The language and rules provide little structure: there are no locutions for to allow for the retraction of previous assertions, no rules to avoid disruptive behaviour and no rules for the terminating the dialogue.

- The semantics are difficult to verify, i.e., there is no way to check the sincerity condition requiring that an agent actually believes a proposition that it states.

The problems associated with FIPA ACL are discussed in more detail by McBurney *et al.* in [109]. The authors also go one step further by providing a list of desiderata for the formal design and assessment of agent argumentation protocols. They discuss to what extent the FIPA ACL, along with other proposals for agent communication protocols, satisfy these desiderata. A more specific discussion of the problems associated with FIPA ACL's semantics is given by Pitt and Mamdani in [124].

Given the problems highlighted above with the attempts to produce a standardised agent communication language, a number of proposals have subsequently been made to address more specific types of communicative agent interaction. In particular, a number of different proposals have been given for agent interaction protocols, including those intended to model the individual types of dialogue of the Walton and Krabbe typology (which was discussed previously in Section 2.4.4). Some examples of such protocols are: McBurney and Parson's protocol for scientific inquiries [106], Amgoud *et al.'s* negotiation dialogue [3] (which draws upon MacKenzie's dialogue game DC, as discussed earlier in this section) and Dignum *et al.'s* persuasion protocol to enable agents to collectively create intentions [51]. McBurney *et al.* discuss all three of these protocols and evaluate them against the desiderata for argumentative agent communication protocols given in [109]. In the ACL literature numerous other proposals for protocols have been given to deal with a wide range of dialogue types: those that been classified into typologies, other distinct types of dialectical interactions, e.g., [110], and even dialogues where one type is embedded within another [107]. Many of these protocols do overcome some of the problems highlighted above with the FIPA ACL, e.g., by providing locutions to question and contest information, and by providing rules to discourage disruptive behaviour. However, the development of such context specific protocols does detract from one of the main aims of the FIPA ACL: to have a standard language for communication amongst agents. Nonetheless, the insights gained from the representation of such distinct dialogical exchanges do contribute to our understanding of the components needed to provide a language expressive enough to represent all of these interactions. The area of agent communication is still relatively young and is currently receiving considerable attention from the multi-agent systems community. In 1999 a review of the then current state-of-the-art within the field was given in [98]. A more recent survey of agent communication languages and protocols, and the different approaches used for their development, is given by Maudet and Chaib-Draa in [104].

The dialogue game protocol proposed in this thesis is based upon the theory of persuasion over action, which will be described in Chapter 3, and the protocol embodying this theory is then set out in Chapter 4. As this protocol derives from a particular argument structure i.e., a new and unique form of argument scheme with associated critical questions, it contains moves that are not covered by other existing persuasion protocols. This is not meant to detract from the worth of these other protocols but it is intended to contribute to a more fine grained representation of the all elements that need to be considered when conducting dialogues about actions to be taken. This concludes my discussion of agent communication languages and protocols.

2.7 Summary and Conclusions

In this chapter I have presented an overview of a number of areas that I have drawn upon for the work I present in the forthcoming chapters. The main background setting for my research is the area of practical reasoning. I have discussed some of the important features of such reasoning and highlighted some of the problems inherent in it and its computational representation. I have also discussed the philosophical topic of argumentation theory and how this subject has proved to be of great use in its application to agent technologies. Finally, I have discussed the topic of multi-agent communication and the current issues and developments within this field.

The prior research presented in this literature survey provides numerous key points that will be taken forward by the work I present in the forthcoming chapters of this thesis. I summarise these key points below.

Practical Reasoning. The account of practical reasoning presented in this thesis is intended to accommodate numerous distinct features of practical reasoning, as identified in the philosophy literature. Firstly, accounts of practical reasoning in philosophy highlight a number of important points: practical reasoning is undertaken in the context of a debate which incorporates a set of arguments for and against the adoption of some particular action; rational disagreement should be accounted for as people have different interests and values which mean that they may not always be able to reach agreement; preferences emerge from debates rather than acting as input to them; theories based upon expected returns can appear counter-intuitive; and, it can be argued that there can be no deductive logic of practical reasoning. Searle [146] advocates all these points and it is broadly his account of practical reasoning that I shall follow. Secondly, from the philosophy and jurisprudence literature, Perelman gives us the notion of an audience. Audiences are differentiated according to their values and the relative priorities they ascribe to these values. So, the acceptability of an argument depends, at least in part, on the audience to which it is being addressed. The model of practical reasoning that I articulate in this thesis is intended to account for these important features of practical reasoning, in order to provide a realistic computational account of such reasoning. Chapters 3, 4 and 5 articulate and develop this model.

Autonomous Agents and Multi-Agent Systems. The aim of this thesis is to provide an account of practical reasoning that can be deployed in autonomous agents. The particular computational setting that will used is the Belief-Desire-Intention (BDI) model of agents. The process by which agents choose what to do is known as ‘the Deliberation Process’. This comprises two parts: the generation of candidate options and the filtering of these options to determine what intentions the agent should commit to. Chapter 5 will demonstrate how this deliberation process for autonomous agents is effected within the account I propose.

Argumentation Theory. The account of argument that will be used in this thesis follows that of Walton [164] who has given a proposal for treating argument as presumptive justification subject to critical questioning. This is manifest through the notion of argument schemes and characteristic critical questions. In particular, Walton's sufficient condition scheme for practical reasoning is of most relevance. This scheme and its critical questions will be examined in greater detail and a proposal for its extension will also be given in Chapter 3.

Argumentation in AI. Argumentation theory provides a number of mechanisms which are useful in their application to AI. One particularly important contribution is Dung's notion of an 'argumentation framework' for the evaluation of the dialectical status of arguments. This mechanism enables us to assess how well an argument on a particular matter can defend itself against attack from other arguments. Dung's framework is extended by Bench-Capon to include the notions of audiences and values and this type of framework will be used in the working examples presented in Chapters 6, 7 and 8.

Agent Communication and Interaction Communication languages for autonomous agents are an integral requirement for the exchange information between autonomous agents. In order for structured and meaningful dialogue to take place, a number of proposals have been given for dialogue game protocols that are designed to facilitate the conduct of particular types of dialogue, such as negotiation dialogues, deliberation dialogues, etc. Additionally, a number of these protocols are based upon dialogue games that have been previously specified in the philosophy literature and thus we are able to represent explicit theories of interaction through these protocols. In Chapter 4 I specify one such dialogue game protocol which adheres to the theory of persuasion over action detailed in Chapter 3. This protocol is intended to enable agents to exchange arguments and information in accordance with my theory.

Each of these areas plays an important part in the proposals I present in the impending chapters for the computational representation of persuasive argument in practical reasoning.

Chapter 3

Theory of Persuasion Over Action

In this chapter I will present an argument scheme for practical reasoning which extends Walton's *sufficient condition scheme* for practical reasoning that was discussed in Section 2.4.3. In Section 3.1 of this chapter I discuss some issues associated with Walton's scheme and I propose an extension to it to address these issues, along with an extended list of critical questions. In Section 3.2 I take the informal description of the extended argument scheme and give it a more precise definition to enable its formal representation. Section 3.3 does the same for the critical questions associated with the scheme, turning them in to attack relations on the proposal for action so that they may later be represented as argumentation frameworks. Section 3.4 gives a summary of the theory detailed in this chapter which will form the basis for a dialogue game protocol named the *PARMA Action Persuasion Protocol*, as will be discussed in Chapter 4.

3.1 Extension of Walton's Scheme for Practical Reasoning

In Section 2.4.2 I discussed the treatment of practical reasoning through the use of argument schemes and critical questions. Here I return to one of the schemes for practical reasoning given by Walton and first discuss some particular issues associated with this scheme, before proposing an extension to it.

Recall from Chapter 2, Walton gives two schemes for practical reasoning in [164]: the *necessary condition scheme* (called W1):[1]

[1] In this and the next scheme, I label each of Walton's symbols for clarity. As noted in the previous chapter, Walton gave a more detailed version of these schemes in an earlier account [163]. I have built this work on the later account.

W1 G is a goal for agent *a*
Doing action A is necessary for agent *a* to carry out goal G
Therefore agent *a* ought to do action A.

and the *sufficient condition scheme* (W2):

W2 G is a goal for agent *a*
Doing action A is sufficient for agent *a* to carry out goal G
Therefore agent *a* ought to do action A.

Walton associates four critical questions with these schemes:

CQW1 Are there alternative ways of realising goal G?
CQW2 Is it possible to do action A?
CQW3 Does agent *a* have goals other than G which should be taken into account?
CQW4 Are there other consequences of doing action A which should be taken into account?

Here I will consider only W2: W1 is a special case in which CQW1 is answered in the negative. The argument scheme given in W2 and the critical questions can both be elaborated because the notion of a goal is overloaded, potentially referring to any of the direct results of the action, consequences of those results, and the reasons why those consequences are desired. These distinctions are of potential importance when considering how to overcome the problems associated with the practical syllogism that were highlighted in Section 2.1.3. Consider the following situation. I am in Liverpool. My friend X is currently in London (200 miles distant) and is about to go to Australia indefinitely. I am eager to say farewell to him. To catch him before he leaves London, it is necessary that I arrive in London before 4.30 pm. As previously discussed in Chapter 2, practical reasoning is situated and it is therefore important to know the story behind the situation in order to be able to consider all the alternatives available in the particular context. So I may say:

WS1 I want to be in London before 4.30 pm.
The 1.30 pm train arrives in London at 4.15 pm.
So, I shall catch the 1.30 pm train.

Here I am justifying my action in terms of one of its consequences. Alternatively I may say:

WS2 I want to see person X before he leaves London.
The 1.30 pm train arrives in London at 4.15 pm.
So, I shall catch the 1.30 pm train.

Here the action is not justified by its direct consequences, but by something else that follows from them. I do not really desire to be in London at all, except in so far as it is a means to the end of seeing X before he departs for Australia. Alternatively there is a third justification:

WS3 Friendship requires that I see person X before he leaves London.
The 1.30 pm train arrives in London at 4.15 pm.
So, I shall catch the 1.30 pm train.

Here I justify my action not in terms of its direct consequences, nor in terms of a state of affairs which will result from the action, but in terms of the underlying social value which I hope to promote by performing the action.

Thus, I have taken Walton's notion of a goal and separated it into three distinct elements: states, goals and values. In my model I define *states* to be a set of propositions about the world to which we can assign a truth value, *goals* are propositional formulae on this set of propositions, and *values* are functions on goals. The distinction between states and goals is made to represent the important difference between effects of actions which the agent wishes to attain, and the effects which follow from an action but are not necessarily desired by the agent. This fits with Searle's observations detailed in Section 2.2.2 that agents should not need to commit to intending to achieve all of the consequences of actions, as there may be undesirable consequences in addition to the favourable ones. Looking to values, these in turn are different from goals as they provide the actual reasons for which an agent wishes to achieve a goal. Thus, values, as discussed previously in Chapter 2, are viewed as being distinct from goals and not just sub or super goals. This has implications when designing autonomous software agents that can make use of values as it extends their notion of autonomy. By associating states of affairs with values we allow agents to try and realise these values, rather than specific states of affairs which have no given reason as to why they are desirable. Moreover, values relate states of affairs, since a given state of affairs may be desirable through promoting several values, and a given value can be promoted by several states of affairs.

The distinction between the different aspects described above is an important factor in practical reasoning situations where the precise points of contention on an issue should be distinguishable and identifiable. This provides the motivation for the extension of Walton's scheme in this manner.

In general, instead of Walton's

W1a G is a goal for agent *a*

we may write:

P1 Agent *a* wishes to achieve state S so as to bring about goal G which promotes value V.

Note that the answers to CQW1 are different in the cases WS1-3:

- In the case of WS1, I must propose other ways of arriving in London on time, perhaps by driving;
- In the case of WS2 I need not go to London at all; for example, I could drive to Heathrow Airport and say goodbye there;
- In the case of WS3 I need not meet with person X at all; perhaps a telephone call and an apology will be enough to promote friendship.

Given this more refined notion of a goal we can extend CQW1 to:

CQ1a Are there alternative ways of realising the same consequences?
CQ1b Are there alternative ways of realising the same goal?
CQ1c Are there alternative ways of promoting the same value?

CQW2 remains unchanged, but CQW3 can also be elaborated, in that it may be that doing action A realises some other goal which promotes some other value, or it may be that doing A prevents some other goal from being realised:

CQ3a Does doing action A realise some other goal which promotes some other value?
CQ3b Does doing action A preclude some other action which would promote some other value?

Also, CQW4 has two aspects:

CQ4a Does doing action A have a side effect which demotes the value?
CQ4b Does doing action A have a side effect which demotes some other value?

Now that the goal has been divided into these three different elements I can now propose an extended argument scheme which incorporates P1 and makes the factual context explicit:

AS1 In the circumstances R,
we should perform action A,
to achieve new circumstances S,
which will realise some goal G,
which will promote some value V.

Apart from the possibility of the action, Walton does not consider other problems with soundness of W2, presupposing that the second premise is to be understood in terms of what agent *a* knows or reasonably believes.

It could be that:

- Action A is not sufficient to bring about goal G; either because the current circumstances are not as presupposed, or because, although the beliefs about the current situation are correct, action A does not have the believed effects.
- Goal G is not a goal for agent *a*; either because there is some problem with the link between the circumstances brought about by doing action A with the value agent *a* assumes them to promote, or because goal G is not in fact a possible state of affairs.

The following critical questions can therefore be added:

CQ5 Are the circumstances such that doing action A will bring about goal G?
CQ6 Does goal G promote value V?
CQ7 Is goal G possible?

Note that an answer to CQ5 needs to address four issues:

a) Whether the believed circumstances R are possible.
b) Whether the believed circumstances R are true.
c) Assuming both of these, whether the action A has the stated consequences S.
d) Assuming all of these, whether the action A will bring about the desired goal G.

Similarly, taking the more articulated view of G expressed as P1, means CQ6 needs to address both:

a) Whether goal G does realise the value intended; and
b) Whether the value proposed is indeed a legitimate value.

Also, taking G in terms of P1, CQ7 needs to address both:

a) Whether the situation S believed by agent *a* to result from doing action A is a possible state of affairs
b) Whether the particular aspects of situation S represented by G are possible.

Thus, I now have an elaborated set of critical questions: four variants of CQ5; three variants of CQ1; two variants of each of CQ3, CQ4, CQ6 and CQ7; and CQ2, making sixteen questions in all. I will use these sixteen critical questions as the basis for the development of my general theory of persuasion over action. Firstly, I re-number the questions into a more logical order that reflects the structure of the argument scheme they are associated with (AS1) and I will use this ordering throughout the rest of the thesis:

CQ1: Are the believed circumstances true?
CQ2: Assuming the circumstances, does the action have the stated consequences?
CQ3: Assuming the circumstances and that the action has the stated consequences, will the action bring about the desired goal?
CQ4: Does the goal realise the value stated?
CQ5: Are there alternative ways of realising the same consequences?
CQ6: Are there alternative ways of realising the same goal?
CQ7: Are there alternative ways of promoting the same value?
CQ8: Does doing the action have a side effect which demotes the value?
CQ9: Does doing the action have a side effect which demotes some other value?
CQ10: Does doing the action promote some other value?
CQ11: Does doing the action preclude some other action which would promote some other value?
CQ12: Are the circumstances as described possible?
CQ13: Is the action possible?
CQ14: Are the consequences as described possible?
CQ15: Can the desired goal be realised?
CQ16: Is the value indeed a legitimate value?

To summarise, in an argument about a matter of practical action, we should expect to see one or more *prima facie* justifications advanced stating, explicitly or implicitly, the current situation, an action, the situation envisaged to result from the action, the features of that situation for which the action was performed, and the value promoted by the action. The critical questions can then be used to generate families of attacks on these justifications. In the next subsections I will give formal definitions for these attacks.

The theory presented here is intended to aid the process of practical reasoning, as described in Chapter 2. The specific situation being considered is where one agent is attempting to persuade another to adopt a course of action, and that other agent is possibly arguing against this. Because this situation is seen as one of conflict, I will refer to the various critical questions as ‘attacks’. Persuasion is intended to be rational, and so reasons are advanced, attacked and defended by each side. This form of persuasion is intended to lead to action. I now present a formal definition of the statement for the justification for an action, as specified by argument scheme AS1, and the sixteen different ways in which it can be attacked, according to the sixteen critical questions associated with AS1.

3.2 Stating a Position

I now present a more formal definition for AS1. However, it should be first noted that no difference needs to be recognised between deciding on a future action and justifying a past action. Moreover, an action may achieve multiple goals, and each goal may promote multiple values. For simplicity, the assumption is made that the proponent of an action articulates an argument in the form of scheme AS1 for each goal realised and value promoted. Thus, the scheme may then be formalised as follows. Assume the existence of:

- A finite set of distinct actions[2], denoted *Acts*, with elements, A, B, C, etc.
- A finite set of propositions, denoted *Props*, with elements, p, q, r, etc.
- A finite set of states, denoted *States*, with elements, R, S, T, etc. Each element of *States* is an assignment of a truth value from the set $\{T, F\}$ to every element of *Props*.
- A finite set of propositional formulae, *Goals*, called goals, with elements G, H, etc.
- A finite set of values *Values*, with elements v, w, etc.
- A function *value* mapping each element of *Goals* to a pair $< v, sign >$, where $v \in$ *Values* and $sign \in \{+, =, -\}$.
- A ternary relation *apply* on *Acts* × *States* × *States*, with *apply(A, R, S)* to be read as: “*Performing action A in state R results in state S.*”

The argument scheme AS1 contains reference to actions and deontic modalities which are not readily formalised in classical logic. We can, however, see that there are four statements of classical logic which must hold if the argument represented by scheme AS1 is to be valid:

[2] By action I do not mean atomic action: an action can cover any coherent set of actions, e.g., “going to London” is an action. The granularity of actions depends on the context. In some cases actions may require highly complex plans, e.g., “declaring war on a country.”

Statement 1: R is the case.

Statement 2: *apply(A, R, S)* $\in$ *apply*.

Statement 3: S $\models$ G (G is true in state S).

Statement 4: *value(G)* $=< v, + >$.

3.3 Attacking a Position

In this subsection I will describe the attacks corresponding to the critical questions presented in Section 3.1. The descriptions are given in terms of the elements identified in the previous subsection, and for each attack I give the source critical question from which it is derived.

I will also consider a number of variants on the basic attacks. When an element of a position is disputed, the attacker may simply disagree, or may additionally offer extra information which indicates the source of the disagreement or makes the disagreement more concrete. Thus, for example, if there is a disagreement as to what is in fact the current situation, an opponent may simply deny what the proponent has said, or may also add what he or she thinks is really the case.

3.3.1 Denial of Premises

A proposal for a particular action A can first be attacked by denying one of the four statements which must obtain for the proposal to be valid. Three of these premises relate to the action realising the goal, and so relate to Critical Questions CQ1, CQ2 and CQ3. The final one concerns the realisation of the claimed value and so relates to CQ4.

Attack 1 (CQ1): R is not the case.

Attack 2 (CQ2): It is not the case that *apply(A, R, S)* $\in$ *apply*.

Attack 3 (CQ3): It is not the case that S $\models$ G.

Attack 4 (CQ4): It is not the case that *value(G)* $=< v, + >$.

Each of these attacks may be executed with differing degrees of force, depending on whether positive information accompanies the attack, and the severity of the consequences of disagreement, and so we are able to distinguish variants of the main attack. Consideration of later elements presupposes agreement on earlier elements of a position for a proposal for action. For example, unless there is agreement on the current circumstances, the effects of an action will not be considered.

Two variant attacks can be identified for **Attack 1**:

Attack 1a: R is not the case.

Attack 1b: R is not the case, and there is a circumstance Q $\in$ States, where R $\neq$ Q, such that Q is the case.

Seven variant attacks can be identified for **Attack 2**:

Attack 2a: It is not the case that *apply(A, R, S)* $\in$ *apply*.

Attack 2b: It is not the case that *apply(A, R, S)* $\in$ *apply*, and it is the case that *apply(A, R, T)* $\in$ *apply*, where T $\neq$ S.

Attack 2c: It is not the case that *apply(A, R, S)* $\in$ *apply*, and it is the case that *apply(A, R, T)* $\in$ *apply*, where T $\neq$ S, but it is not the case that T $\models$ G.

Attack 2d: It is not the case that *apply(A, R, S)* $\in$ *apply*, and it is the case that *apply(A, R, T)* $\in$ *apply*, where T $\neq$ S, and it is the case that T $\models$ H[3], but it is not the case that *value(H)* $=< v, + >$.

Attack 2e: It is not the case that *apply(A, R, S)* $\in$ *apply*, and it is the case that *apply(A, R, T)* $\in$ *apply*, where T $\neq$ S, and it is the case that T $\models$ H, but *value(H)* $=< v, - >$.

Attack 2f: It is not the case that *apply(A, R, S)* $\in$ *apply*, and it is the case that *apply(A, R, T)* $\in$ *apply*, where T $\neq$ S, and it is the case that T $\models$ H, but *value(H)* $=< w, + >$, where w $\neq$ v.

Attack 2g: It is not the case that *apply(A, R, S)* $\in$ *apply*, and it is the case that *apply(A, R, T)* $\in$ *apply*, where T $\neq$ S, and it is the case that T $\models$ H, but *value(H)* $=< w, - >$, where w $\neq$ v.

Similarly, we may distinguish six variants of **Attack 3**:

Attack 3a: It is not the case that S $\models$ G.

Attack 3b: It is not the case that S $\models$ G and there is a goal H $\in$ *Goals*, H $\neq$ G, such that S $\models$ H.

Attack 3c: It is not the case that S $\models$ G and there is a goal H $\in$ *Goals*, H $\neq$ G, such that S $\models$ H and with *value(H)* $\neq< v, + >$.

Attack 3d: It is not the case that S $\models$ G and there is a goal H $\in$ *Goals*, H $\neq$ G, such that S $\models$ H and with *value(H)* $=< v, - >$.

[3] In attacks 2d–2g H may or may not be distinct from G.

Attack 3e: It is not the case that S $\models$ G and there is a goal H $\in$ *Goals*, H $\neq$ G, and a value w $\in$ *Values*, w $\neq$ v, such that S $\models$ H and with *value(H)* $=< w, + >$.

Attack 3f: It is not the case that S $\models$ G and there is a goal H $\in$ *Goals*, H $\neq$ G, and a value w $\in$ *Values*, w $\neq$ v, such that S $\models$ H and with *value(H)* $=< w, - >$.

Likewise, we may distinguish four variants of **Attack 4**:

Attack 4a: It is not the case that *value(G)* $=< v, + >$.

Attack 4b: It is not the case that *value(G)* $=< v, + >$ and *value(G)* $=< v, - >$.

Attack 4c: It is not the case that *value(G)* $=< v, + >$ and there is a value w $\in$ *Values*, w $\neq$ v, such that *value(G)* $=< w, + >$.

Attack 4d: It is not the case that *value(G)* $=< v, + >$ and there is a value w $\in$ *Values*, w $\neq$ v, such that *value(G)* $=< w, - >$.

3.3.2 Alternative Ways to Satisfy the Same Value

These four attacks relate to Critical Questions CQ5, CQ6 and CQ7. They each propose an alternative way of achieving the same desired value.

Attack 5 (CQ5): There exists an action B $\in$ *Acts*, with B $\neq$ A, and *apply(B,R,S)* $\in$ *apply*.

Attack 6 (CQ6): There exists an action B $\in$ *Acts*, with B $\neq$ A, and *apply(B,R,T)* $\in$ *apply*, with T $\models$ G.

Attack 7a (CQ7): There exists an action B $\in$ *Acts*, with B $\neq$ A, and *apply(B,R,T)* $\in$ *apply*, with T $\models$ H, and *value(H)* $=< v, + >$.

Attack 7b (CQ7): There is a goal H $\in$ *Goals*, with H $\neq$ G, such that *apply(A,R,S)* $\in$ *apply* with S $\models$ H, and with *value(H)* $=< v, + >$.

3.3.3 Side Effects of the Action

Two of these attacks relate to unconsidered consequences of the action, raised by Critical Questions CQ8 and CQ9. The third offers a different justification for the action, and so relates to other goals that need to be considered, as in Critical Question CQ10.

Attack 8 (CQ8): There is a goal H $\in$ *Goals*, with H $\neq$ G, such that *apply(A,R,S)* $\in$ *apply* with S $\models$ H, and with *value(H)* $=< v, - >$.

Attack 9 (CQ9): There is a goal H $\in$ *Goals*, with H $\neq$ G, and there is a value w $\in$ *values*, with w $\neq$ v, such that *apply(A,R,S)* $\in$ *apply* with S $\models$ H, and with *value(H)* $=< w, - >$.

Attack 10 (CQ10): There is a goal H $\in$ *Goals*, with H $\neq$ G, and there is a value w $\in$ *values*, with w $\neq$ v, such that *apply(A,R,S)* $\in$ *apply* with S $\models$ H, and with *value(H)* $=< w, + >$[4].

3.3.4 Interference with Other Actions

This group of attacks all relate to the promotion of some other value, and so derive from Critical Question CQ11. The three variants arise respectively from: consideration of the compatibility of the proposed action with some other action; whether the proposed action realises a state of affairs incompatible with the goal of another action; or whether the state of affairs realised is incompatible with *all* ways of promoting some other value.

Attack 11a: It is the case that *apply(A,R,S)* $\in$ *apply*. There is a value w $\in$ *values* with w $\neq$ v. There is an action B $\in$ *Acts* with B $\neq$ A, such that *apply(B,R,T)* $\in$ *apply*, with T $\models$ H, and *value(H)* $=< w, + >$. However, there is no state X $\in$ *States* such that *apply(A&B,R,X)*[5] $\in$ *apply*.

Attack 11b: It is the case that *apply(A,R,S)* $\in$ *apply*. There is a value w $\in$ *values* with w $\neq$ v. There is a goal H $\in$ *Goals* with H $\neq$ G, such that *value(H)* $=< w, + >$. However, S $\models$ $\neg$H.

Attack 11c: It is the case that *apply(A,R,S)* $\in$ *apply*. There is a value w $\in$ *values* with w $\neq$ v. However, if there is a goal J $\in$ *Goals*, with *value(J)* $=< w, + >$, then S $\models$ $\neg$J.

3.3.5 Disagreements Relating to Impossibility

The final group of attacks all relate to whether an element of the position is possible or not. In the critical questions I considered possibility together with the other questions relating to the element under dispute. Therefore these attacks relate to a number of different critical questions, as indicated below.

Attack 12 (CQ12): It is not the case that R $\in$ *States*.

Attack 13 (CQ13): It is not the case that A $\in$ *Acts*.

Attack 14 (CQ14): It is not the case that S $\in$ *States*.

[4] Note that Attacks 7b and 10 do of themselves dispute that the action should be performed, nor that a value will be promoted. Their significance comes when the discussion concerns the motivation for performing a particular action. Such attacks becomes important where justification of a past action is taken as a precedent for some future action (for example, in legal applications), and in situations where arguments lend support to one another, as will be noted in the examples of Chapters 6 and 8.

[5] A&B denotes the execution of two actions, A and B, which could be conducted sequentially or in parallel.

Attack 15 (CQ15): It is not the case that G $\in$ *Goals*.

Attack 16 (CQ16): It is not the case that v $\in$ *Values*.

A summary of these attacks and the number of variants related to each one can be seen in Table 3.1. Additionally, the last column of the table indicates the nature of the dispute manifest in the attack and this provides a basis for resolution of the conflict. Responses to attacks and resolution of conflicts will be discussed in Chapter 4.

Table 3.1: **Attacks on a Proposal for Action**

Attack	Variants	Description	Basis of Resolution
1	2	Disagree with the description of the current situation	What is true
2	7	Disagree with the consequences of the proposed action	What is true
3	6	Disagree that the desired features are part of the consequences	Representation
4	4	Disagree that these features promote the desired value	What is true
5	1	Believe the consequences can be realised by some alternative action	What is best
6	1	Believe the desired features can be realised through some alternative action	What is best
7	2	Believe that the desired value can be realised in an alternative way	What is best
8	1	Believe the action has undesirable side effects which demote the desired value	What is best
9	1	Believe the action has undesirable side effects which demote some other value	What is best
10	1	Agree that the action should be performed, but for different reasons	What is best
11	3	Believe that the action will preclude some more desirable action	What is best
12	1	Believe that the circumstances as described are not possible	Representation
13	1	Believe that the action is impossible	What is true
14	1	Believe that the consequences as described are not possible	Representation
15	1	Believe that the desired features cannot be realised	Representation
16	1	Disagree that the desired value is a legitimate value	Representation

3.4 Summary

In this chapter I have discussed an argument scheme and accompanying critical questions for practical reasoning that has been proposed by Walton. I drew attention to some issues associated with this scheme and proposed an extension to it to deal with these issues. This extended scheme and critical questions are intended to enable subjective components of practical reasoning to be accounted for. Representation of these subjective components follows Searle's intuitions, detailed Chapter 2, which place importance upon the recognition of subjective elements of practical reasoning. I followed my proposal of the extended argument scheme and critical questions with a set of more formal definitions. The definitions set out here can now be specified in terms that will enable them to be used as the basis for a dialogue game protocol, where participants can put forward and attack proposals for action, in accordance with the theory given in this chapter. I articulate this dialogue game protocol in the next chapter.

Chapter 4

The PARMA Protocol

Chapter 3 presented a theory which lays the foundations for a multi-agent dialogue game protocol. In this chapter I will articulate this protocol, which I call *PARMA* (for Persuasive ARgument for Multiple Agents) *Action Persuasion Protocol*, in more detail. This protocol will enable persuasive argument over proposed courses of action to be undertaken by two or more participants of a dialogue. A proponent of an action may state and justify his or her proposal for action in the form of argument scheme AS1, and opponents may attack this position according to the 16 different attacks (and their variants) presented in Chapter 3. Section 4.1 presents the syntax of PARMA. Section 4.2 outlines an axiomatic semantics for the protocol in terms of pre- and post-conditions on the participants' commitment stores. Section 4.3 provides a discussion of the different categories that each of the attacks falls under and how resolution of conflicts is dependant upon the category of the attack. Section 4.4 describes a system that implements the PARMA Protocol to mediate dialogues about actions between human participants. Here a description and evaluation of the implemented system are given. Section 4.5 concludes with a summary.

4.1 Syntax of PARMA

In this section I present the syntax of the PARMA Protocol. Following the discussion on agent communication from Section 2.6.2, I assume that the language syntax comprises two layers: an inner layer in which the topics of conversation are represented formally, and an outer, or wrapper, layer comprising locutions which express the illocutionary force of the inner content. In the presentation of the axiomatic semantics propositional logic is assumed as the formal representation of the inner layer, but this restriction is for simplicity of presentation only.

The syntax of PARMA is presented by listing the twenty-five legal locutions in Tables 4.1 and 4.2, grouped into five classes. Fifteen locutions are shown in Table 4.1, grouped into three classes (columns): locutions to control the dialogue; locutions to ask about an agent's position; and locutions to state a position for the justification of an action.

Table 4.1: **Locutions to control the dialogue, ask about a position and state a position**

'Control' Locutions	'Ask' Locutions	'State' Locutions
Enter dialogue	Ask circumstances(R)	State circumstances(R)
Leave dialogue	Ask action(A)	State action(A)
Turn finished	Ask consequences(A,R,S)	State consequences(A,R,S)
Accept denial	Ask logical consequences(S,G)	State logical consequences(S,G)
Reject denial	Ask purpose(G,V,D)	State purpose(G,V,D)

Table 4.2 contains another ten locutions, grouped into two classes (columns): locutions to attack elements of a position; and locutions to attack the validity of elements of a position.

Table 4.2: **Locutions to attack a position and attack the validity of elements**

'Deny' Locutions	'Deny Existence' Locutions
Deny circumstances(R)	Deny initial circumstances exist(R)
Deny action(A)	Deny action exists(A)
Deny consequences(A,R,S)	Deny resultant state exists(S)
Deny logical consequences(S,G)	Deny goal exists(G)
Deny purpose(G,V,D)	Deny value exists(V)

4.2 Axiomatic Semantics of PARMA

I now present an axiomatic semantics for the PARMA Protocol. A point to note regarding the axiomatic semantics specified here is that the set of conditions that follows is just one of a number of possible alternative specifications founded upon the general theory of persuasion over action that was detailed in the previous chapter. The set of conditions given here somewhat strictly enforce the protocol, whereas it would equally be possible to give a rather 'looser' version of the semantics to allow for greater flexibility in the protocol, if this was deemed necessary for a particular domain.

4.2.1 Axiomatic Semantics

Tables 4.3–4.7 present the pre-conditions necessary for the legal utterance of each locution under the protocol, and any post-conditions arising from their legal utterance. Thus, Tables 4.3–4.7 present an outline of an axiomatic semantics [156, 58] for the

PARMA Protocol, and imply the rules governing the combination of locutions under the protocol [107]. It is assumed, following [80] and as discussed in Section 2.6.1, that a *commitment store* is associated with each participant. Each commitment store records, in a manner which all participants may read, the commitments made by that participant in the course of a dialogue. It is also assumed that a history of the dialogue is available to all participants. The post-conditions of utterances shown in Tables 4.3–4.7 include any commitments incurred by the speaker of each utterance while the pre-conditions indicate any prior commitments required before an utterance can be legally made. Commitments in this protocol are dialogical and follow Hamblin's notion of commitment – i.e., statements which an agent must defend if attacked, and may bear no relation to the agent's real beliefs or intentions [80]. Once a move has legally been executed by a player, the turn can be passed, where the next player then has a set of moves from which the choice of the next utterance may be made. These 'next available moves' are entirely defined by the pre-conditions of the locutions.

Section 4.4 will provide a short description and evaluation of an implemented dialogue game system built upon the semantics that follow.

Table 4.3: **Locutions to control the dialogue**

Locution	Pre-conditions	Post-conditions
Enter dialogue	Speaker has not already uttered enter dialogue	Speaker has entered dialogue
Leave dialogue	Speaker has uttered enter dialogue	Speaker has left dialogue
Turn finished	Speaker has finished making their move	Speaker and hearer switch roles so new speaker can now make a move
Accept denial	Hearer has made an attack on an element of speaker's position	Speaker committed to the negation of the element that was denied by the hearer
Reject denial	Hearer has made an attack on an element of speaker's position	Point of disagreement reached

Table 4.4: **Locutions to propose an action**

Locution	Pre-conditions	Post-conditions
State circumstances(R)	Speaker uttered enter dialogue. Hearer uttered enter dialogue.	Speaker committed to R. Speaker committed to R $\in$ States.
State action(A)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Speaker committed to R. Speaker committed to R $\in$ States.	Speaker committed to A. Speaker committed to A $\in$ Acts.
State consequences(A,R,S)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Speaker committed to R. Speaker committed to R $\in$ States. Speaker committed to A. Speaker committed to A $\in$ Acts.	Speaker committed to apply(A,R,S) $\in$ apply. Speaker committed to S $\in$ States.
State logical consequences(S,G)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Speaker committed to R. Speaker committed to R $\in$ States. Speaker committed to A. Speaker committed to A $\in$ Acts. Speaker committed to apply(A,R,S) $\in$ apply. Speaker committed to S $\in$ States.	Speaker committed to S $\models$ G. Speaker committed to G $\in$ Goals.
State purpose(G,V,D)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Speaker committed to R. Speaker committed to R $\in$ States. Speaker committed to A. Speaker committed to A $\in$ Acts. Speaker committed to apply(A,R,S) $\in$ apply. Speaker committed to S $\in$ States. Speaker committed to S $\models$ G. Speaker committed to G $\in$ Goals.	Speaker committed to (G,V,D). Speaker committed to V $\in$ Values.

Table 4.5: **Locutions to ask about an agent's position**

Locution	Pre-conditions	Post-conditions
Ask circumstances(R)	Hearer uttered enter dialogue. Speaker uttered enter dialogue. Speaker not committed to circumstances(R) about topic in question.	Hearer must reply with state circumstances(R) or don't know(R).
Ask action(A)	Hearer uttered enter dialogue. Speaker uttered enter dialogue. Speaker not committed to action(A) about topic in question.	Hearer must reply with state action(A) or don't know(A).
Ask consequences(A,R,S)	Hearer uttered enter dialogue. Speaker uttered enter dialogue. Speaker not committed to consequences(A,R,S) about topic in question.	Hearer must reply with state consequences(A,R,S) or don't know(A,R,S).
Ask logical consequences(S,G)	Hearer uttered enter dialogue. Speaker uttered enter dialogue. Speaker not committed to logical consequences(S,G) about topic in question.	Hearer must reply with state logical consequences(S,G) or don't know(S,G).
Ask purpose(G,V,D)	Hearer uttered enter dialogue. Speaker uttered enter dialogue. Speaker not committed to purpose(G,V,D) about topic in question.	Hearer must reply with state purpose(G,V,D) or don't know(G,V,D).

Table 4.6: **Locutions to attack elements of a position**

Locution	Pre-conditions	Post-conditions
Deny circumstances(R)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Hearer committed to R. Hearer committed to R $\in$ States.	Speaker committed to deny circumstances(R).
Deny consequences(A,R,S)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Hearer committed to R. Hearer committed to R $\in$ States. Hearer committed to A. Hearer committed to A $\in$ Acts. Hearer committed to apply(A,R,S) $\in$ apply. Hearer committed to S $\in$ States.	Speaker committed to deny consequences(A,R,S) $\in$ apply.
Deny logical consequences(S,G)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Hearer committed to R. Hearer committed to R $\in$ States. Hearer committed to A. Hearer committed to A $\in$ Acts. Hearer committed to apply(A,R,S) $\in$ apply. Hearer committed to S $\in$ States. Hearer committed to S $\models$ G. Hearer committed to G $\in$ Goals.	Speaker committed to deny logical consequences(S,G) S $\models$ G.
Deny purpose(G,V,D)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Hearer committed to R. Hearer committed to R $\in$ States. Hearer committed to A. Hearer committed to A $\in$ Acts. Hearer committed to apply(A,R,S) $\in$ apply. Hearer committed to S $\in$ States. Hearer committed to S $\models$ G. Hearer committed to G $\in$ Goals. Hearer committed to (G,V,D). Hearer committed to V $\in$ Values.	Speaker committed to deny purpose(G,V,D).

Table 4.7: **Locutions to attack validity of elements**

Locution	Pre-conditions	Post-conditions
Deny initial circumstances exist(R)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Hearer committed to R $\in$ States.	Speaker committed to deny initial circumstances exist(R).
Deny action exists(A)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Hearer committed to R. Hearer committed to R $\in$ States. Hearer committed to A $\in$ Acts.	Speaker committed to deny action exists(A).
Deny resultant state exists(S)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Hearer committed to R. Hearer committed to R $\in$ States. Hearer committed to A $\in$ Acts. Hearer committed to S $\in$ States.	Speaker committed to deny resultant state exists(S).
Deny goal exists(G)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Hearer committed to R. Hearer committed to R $\in$ States. Hearer committed to A $\in$ Acts. Hearer committed to S $\in$ States. Hearer committed to G $\in$ Goals.	Speaker committed to deny goal exists(G).
Deny value exists(V)	Speaker uttered enter dialogue. Hearer uttered enter dialogue. Hearer committed to R. Hearer committed to R $\in$ States. Hearer committed to A $\in$ Acts. Hearer committed to S $\in$ States. Hearer committed to G $\in$ Goals. Hearer committed to V $\in$ Values.	Speaker committed to deny value exists(V).

4.2.2 Locutions for the Attacks

The set of attacks presented in Section 3.3 can now be individually described by combining the previously defined locutions of the dialogue game. The attacks are made up of a mixture of the primitive locutions and the order in which the primitive locutions are presented as part of an attack is of no relevance.

Attack 1a: deny circumstances(R).

Attack 1b: state circumstances(Q) AND deny circumstances(R).

Attack 2a: deny consequences(A,R,S).

Attack 2b: state consequences(A,R,T) AND deny consequences(A,R,S).

Attack 2c: state consequences(A,R,T) AND deny consequences(A,R,S) AND deny logical consequences(T,G).

Attack 2d: state consequences(A,R,T) AND state logical consequences(T,H) AND deny purpose(H,V,D+) AND deny consequences(A,R,S).

Attack 2e: state consequences(A,R,T) AND state logical consequences(T,H) AND state purpose(H,V,D-) AND deny consequences(A,R,S).

Attack 2f: state consequences(A,R,T) AND state logical consequences(T,H) AND state purpose(H,W,D+) AND deny consequences(A,R,S).

Attack 2g: state consequences(A,R,T) AND state logical consequences(T,H) AND state purpose(H,W,D-) AND deny consequences(A,R,S).

Attack 3a: deny logical consequences(S,G).

Attack 3b: state logical consequences(S,H) AND deny logical consequences(S,G).

Attack 3c: state logical consequences(S,H) AND state purpose(H,V,D+) AND deny logical consequences(S,G).

Attack 3d: state logical consequences(S,H) AND state purpose(H,V,D-) AND deny logical consequences(S,G).

Attack 3e: state logical consequences(S,H) AND state purpose(H,W,D+) AND deny logical consequences(S,G).

Attack 3f: state logical consequences(S,H) AND state purpose(H,W,D-) AND deny logical consequences(S,G).

Attack 4a: deny purpose(G,V,D+).

Attack 4b: state purpose(G,V,D-) AND deny purpose(G,V,D+).

Attack 4c: state purpose(G,W,D+) AND deny purpose(G,V,D+).

Attack 4d: state purpose(G,W,D-) AND deny purpose(G,V,D+).

Attack 5: state action(B) AND state consequences(B,R,S).

Attack 6: state action(B) AND state consequences(B,R,T) AND state logical consequences(T,G).

Attack 7a: state action(B) AND state consequences(B,R,T) AND state logical consequences(T,H) AND state purpose(H,V,D+).

Attack 7b: state consequences(A,R,S) AND state logical consequences(S,H) AND state purpose(H,V,D+).

Attack 8: state consequences(A,R,S) AND state logical consequences(S,H) AND state purpose(H,V,D-).

Attack 9: state consequences(A,R,S) AND state logical consequences(S,H) AND state purpose(H,W,D-).

Attack 10: state consequences(A,R,S) AND state logical consequences(S,H) AND state purpose(H,W,D+).

Attack 11a: state consequences(A,R,S) AND state action(B) AND state consequences(B,R,T) AND state logical consequences(T,H) AND state purpose(H,W,D+) AND deny consequences(A&B,R,X).

Attack 11b: state consequences(A,R,S) AND state purpose(H,W,D+) state logical consequences(S,¬ H).

Attack 11c: state consequences(A,R,S) AND IF state purpose(J,W,D+) THEN state logical consequences(S,¬ J).

Attack 12: deny initial state exists(R).

Attack 13: deny action exists(A).

Attack 14: deny resultant state exists(S).

Attack 15: deny goal exists(G).

Attack 16: deny value exists(V).

In addition to the axiomatic semantics presented above, a denotational semantics for the PARMA Protocol has also been specified in joint work with McBurney and Bench-Capon. The axiomatic semantics detailed here are sufficient to enable implementation of the protocol and they serve as the main focus of the semantic element of PARMA. Thus, the outline of a denotational semantics for the PARMA Protocol is provided as a supplementary semantics and so it is presented in Appendix A. However, the denotational semantics are a useful addition which provide a means by which we can study the properties of the protocol and reason about dialogues conducted under it. This a branch of work which could be extended and fully specified in future work to gain further insights into the properties of the PARMA Protocol.

4.3 Responding to Attacks

Now that the statement of a position and the criticism of the elements of such a position have been defined, I examine the ways in which the recipient of an attack can respond to their opponent's criticism. The attacks are again recapitulated in Table 4.8.

How a proponent of a proposal for action responds to an attack depends upon the nature of the attack, as shown in Table 4.8. For those attacks which explicitly state an alternative position, the original proponent is able to counter-attack with some subset of the attacks listed in Table 4.8. For example, if a proponent argues for an action on the grounds that this will promote some value v, and an attacker argues in response that the proposed action will also demote some other value w, then the proponent may respond to this attack by arguing that the action does not have this effect on w (Attack 4a), or that an alternative action can promote w (Attack 7a), or that w is not worth promoting (Attack 16), etc. Whether or not two participants may ultimately reach agreement on a proposed action will depend on the relationship between the participants and on the precise nature of the disagreement. Note that resolving certain conflicts may require leaving the persuasion dialogue to enter a dialogue of a different type, described as 'nesting' in [167]. A basis for any resolution between participants for each type of attack is shown in the fourth column of Table 4.8. I will now examine each individual basis for resolution, discussing the precise nature of the dispute and how resolution of the dispute could be reached.

4.3.1 Factual Disagreements

If the disagreement concerns the nature of the current world-state (Attacks 1 and 13), i.e., a dispute about "What is true", then some process of agreed empirical investigation may resolve this difference between the participants. The same process would also apply to the resolution of disputes regarding causal relations (Attacks 2 and 4). This may involve the participants entering a sub-dialogue, perhaps involving a third party outside

Table 4.8: **Attacks on a Proposal for Action**

Attack	Description	Basis of Resolution
1	Disagree with the description of the current situation	What is true
2	Disagree with the consequences of the proposed action	What is true
3	Disagree that the desired features are part of the consequences	Representation
4	Disagree that these features promote the desired value	What is true
5	Believe the consequences can be realised by some alternative action	What is best
6	Believe the desired features can be realised through some alternative action	What is best
7	Believe that the desired value can be realised in an alternative way	What is best
8	Believe the action has undesirable side effects which demote the desired value	What is best
9	Believe the action has undesirable side effects which demote some other value	What is best
10	Agree that the action should be performed, but for different reasons	What is best
11	Believe that the action will preclude some more desirable action	What is best
12	Believe that the circumstances as described are not possible	Represenation
13	Believe that the action is impossible	What is true
14	Believe that the consequences as described are not possible	Represenation
15	Believe that the desired features cannot be realised	Representation
16	Disagree that the desired value is worth promoting	Representation

their own dialogical exchange, in order to resolve the dispute through the elicitation of the authoritative knowledge of the third party. Alternatively one of the participants may have a role in the dialogue which entitles the opinion of that party to be authoritative (cf. [150]).

4.3.2 Different Preferences

Disputes about “What is best” relate to the preferences of the individual participants. Often such disputes arise from participants ranking their preferences differently (cf. the discussions of Searle from Chapter 2 on this matter). Thus, there is no dispute as to the possibility of the performance of, for example, the action in question, but a dispute can arise due to one party believing the action not to be the best one to perform in the given situation. As mentioned in Section 3.3, there may be a number of reasons as to why a participant does not endorse their opponent's action. There may be alternative possible actions which have the same effect of producing the desired results and any such alternative actions may be more preferable to a participant (Attacks 5, 6 and 7). Conversely, an action may have previously unconsidered detrimental side effects, with respect to the goals it achieves and the values promoted by these goals (Attacks 8, 9 and 10). Finally, a participant may deem an action as undesirable if it interferes with other actions in question, with respect to the promotion of another value, previously not considered (Attack 11). In such cases, disputes must be decided by determining the party whose wishes are to be represented, by constructing a preference order [53] or by some form of negotiation.

4.3.3 Representation

Disputes which relate to representation issues are concerned with the language being used and the logic being deployed in the argument (Attacks 3, 12, 14, 15 and 16). Language is intrinsically connected with meaning and understanding; thus, if both parties involved in the dialogue speak the same language and are competent users of an agreed logic, then the resolution of a dispute over representation should be straightforward. One way of ensuring that computer agents share the same language and concepts is through commitment to the same ontologies, to establish the common language of the topic in question. Ontological differences and their resolution are discussed in [33, 154, 155].

The model presented here assumes that such matters of meaning and context are agreed upon by the participants of a dialogue beforehand and therefore such attacks concerning representation should not occur frequently in dialogue exchanges. However, these attacks remain possible, especially in systems which permit encounters with unfamiliar or unpredictable agents, and should not be overlooked.

4.3.4 Clarification of a Position

Disputes are commonly caused in everyday conversations through participants making ill-informed assumptions about each other's positions. As conversations progress the players' positions become clearer and more explicit and earlier ill-informed assumptions may be dissolved. However, players can recognise that they are not aware of their opponent's full position on an issue (and they may not even be aware of their own full position on an issue)[1]. If the position is not fully explicit then the players may have to elucidate their opponent's position through questioning in order to be able to make an attack on it.

[1] This is discussed by Walton and Krabbe in [167] as dark-side commitments.

4.3.5 Resolution

Successful resolution of a dispute partially depends upon which of the above types of dispute is encountered. Disputes over facts should be straightforwardly resolved if some process of empirical investigation is agreed upon between the participants. Issues of representation should also not be difficult to resolve if participants agree on language and context before the dialogue starts, and participants' ontologies are aligned to ensure a shared understanding of the concepts in the given topic of conversation. Both disagreements about representation and disagreements about facts should be resolved before disagreements about choice are addressed.

Resolution of disputes about what is best typically depends on the context in which the dialogue is taking place. It may be the case that one party is an authority on the matter in question and this will facilitate resolution. For example, in government issues it is usual for government advisors to find out the facts of the situation and for ministers to make the choices between possible actions on the basis of these facts, in the light of the ministers' values. The advisors are then authorities as to facts, as the ministers are authorities as to values. Similarly in a court case, juries are authoritative as to facts, while the role of the judge is to choose legal interpretations.

Naturally, resolution will also occur if one party allows himself to be persuaded that his preference ordering is wrong or if he concedes to the ordering of his opponent's preferences. If agents are able to agree on preferences over actions and over values then they should be able to agree overall. However, if the participants disagree over which value should be promoted in the current situation, then resolution may require agreement between them on a preference ordering over values. Such resolution may require other types of dialogue, and some of these interactions have received considerable attention from philosophers, for example [78, 123, 142]. A formalism to represent disagreement involving arguments which rely on values is proposed in [26] and is discussed further in Chapter 5 onwards.

When there is no authority on the matter to whom an appeal can be made, then we must consider *how* the question of what is best is decided. Two phenomena need to be respected: the possibility of rational disagreement, and value preferences emerging from the reasoning. As to rational disagreement, it is simply not the case that everyone need make the same choices. Not only may different agents have different desires, but they also may legitimately take different views on what is best. Recall, as discussed in Chapter 2, both Perelman and Searle give compelling cases to explain why rational agents may disagree on matters due their own subjective values and perspectives. With regard to emerging values, although many current agent systems use a general utility function, Searle believes that preferences are the product and not the input to practical reasoning, as stated in one of his quotes cited in Chapter 2. If Searle is right, and intuitively it seems more plausible than arguing that all people make their selections according to pre-existing utility functions, this issue needs also to be accounted for. Therefore, we need to employ some method for choosing between alternatives. So, after disputes relating to representation and fact have been addressed, we are left with a number of competing arguments to the effect that an action should or should not be performed, each of them deriving their strength from the value they promote or demote. The set of competing arguments suggests that we could use an argumentation framework such as that developed by Dung in [55] to resolve factual disagreements. To accommodate the strength of arguments in terms of values, we can use the extension of this framework to accommodate values developed by Bench-Capon in [26]. How this may be achieved is discussed in detail in Chapter 5. In both [55] and [26], the use of preferred semantics gives rise to the possibility of different but defensible choices, thus accommodating the possibility of rational disagreement. Similar issues have also been explored in systems designed to mediate human to human dialogues, such as [40, 71, 88, 101]. Doutre *et al.* [53] address the issue of preference ordering on values and define a dialogue which allows value preferences to emerge from the dialogue.

To summarise, successful resolution of a dispute depends upon a number of issues including; the type of dispute encountered, the relationship between the participants, and their individual preference orderings. But it must also be noted that the model presented here should and does allow for the possibility of rational disagreement; it is often a difficult task to persuade others to change their ranking of personal values, and thus such arguments could terminate in conflict. Resolution of conflicts may also be achieved by an agreed procedure, such as voting, or the agents may agree to disagree. In summary, where there is no 'right' answer we must always model the possibility of different, but acceptable solutions, upholding the observations made by Searle on this matter, as was discussed in Chapter 2 .

4.4 Implementation of the PARMA Dialogue Game Protocol

I shall now describe an implemented system which takes the form of a dialogue game and embodies the PARMA Protocol articulated in the previous sections of this chapter. This description will encompass a brief outline of the system along with a summary of its merits and shortcomings. The objective of the implementation presented here is to provide a proof of concept for the PARMA Protocol. The implementation also represents a step towards the ultimate goal of allowing persuasive dialogue between autonomous software agents, as will be discussed in Chapter 5. A more detailed description of the implemented system, as well as the accompanying design documentation, can be found in Appendix B and [11].

4.4.1 General Description

I have implemented the *PARMA Action Persuasion Protocol* in the form of a Java program and I will now give a brief description of how the system functions. The program implements the protocol so that dialogues between two human participants can be undertaken under the protocol, with each participant taking turns to propose and attack positions uttering the locutions specified in Section 4.2. The legality of the participants' chosen moves is checked by the program through verification that all pre-conditions for moves hold. Thus, the participants are able to state and attack each other's positions with the program verifying that the dialogue always complies with the protocol. If a participant attempts to make an illegal move then they are informed of this and given the opportunity to chose an alternative move. After a move has been legally uttered, the commitment store of the participant who made the move is updated to contain any new commitments incurred by the utterance. All moves, whether legal or illegal, are entered into the history, which records which moves were made by which participant and the legality of the move chosen. Additionally after a move has been legally uttered, the commitment store of the player who made the move is printed to the screen to show all previous commitments and any new ones that have consequently been added. By publicly displaying the commitment stores in this way each participant is able to see their own and each other's commitments. So, participants can determine which of their commitments overlap with those of the other participant, and thereby identify points of agreement. Conversely, this also allows each participant to identify any commitments of the other participant in conflict with their own, and thus which commitments are susceptible to an attack.

Dialogues undertaken via the program can terminate in a number of ways (see Figure 4.1). A participant can decide to leave the game by exiting at any time, thereby terminating the dialogue. A dialogue can also terminate if disagreement about a position is reached. This occurs when a participant states an element of a position which is subsequently attacked by the other participant, and the first participant disputes the validity of the attack. If the first participant refuses to accept the reasons for the attack then disagreement has been identified and the dialogue terminates. Dialogues may also reach a natural end with agreement between the two participants on a course of action. If this occurs, both players may choose to exit the dialogue. Note that the implemented game addresses only persuasion dialogues. In a system able to handle other types of dialogue, exiting could lead to a dialogue of another type, e.g., an information-seeking dialogue to resolve a factual disagreement.

When a dialogue terminates, whether in agreement or disagreement, the history and commitment stores of both players are printed on screen and also to a file. The dialogue may then be analysed, for example, to see which attacks occurred, or how often or how successful they were. Such analysis may be useful for a study of appropriate strategies for dialogues conducted under the protocol.

4.4.2 State Transition Diagram

Figure 4.1 gives a simple state transition diagram for the dialogue game protocol and it is intended as a high level supplement to the design documentation for the program given in Appendix B. The diagram shows the types of moves that the players can make and the choice of move which is then available in the new state. It also shows the moves that lead to the roles of speaker and hearer being switched and how the game can terminate. The diagram does not show the specific details of all moves that can be made, only the types of moves. For example a ‘state’ move can be made by uttering any of the locutions for proposing an action, given in Table 4.4 of the axiomatic semantics, e.g., ‘state circumstances’, ‘state action’, ‘state consequences’ etc. A ‘deny’ move can be made by uttering any of the locutions for attacking a position, given in Tables 4.6 and 4.7 of the axiomatic semantics, e.g., ‘deny circumstances’, ‘deny consequences’, ‘deny action exists’ etc. An ‘ask’ move can be made by uttering any of the locutions for asking about an agent’s position, given in Table 4.5 of the axiomatic semantics, e.g., ‘ask circumstances’, ‘ask action’, ‘ask consequences’ etc. Which moves of a given type are possible in a particular situation depends on the moves made to reach a state: for example, what can be denied depends on what has been previously asserted.

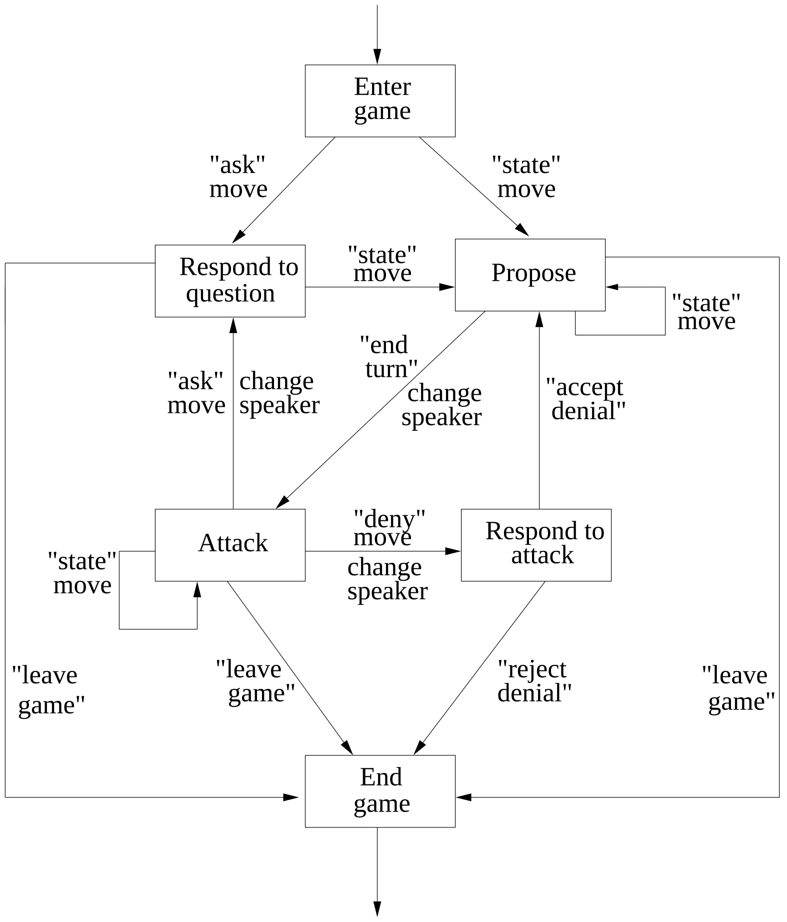


Figure 4.1: State Transition Diagram for *PARMA* Dialogue Game Protocol.

4.4.3 Summary of Issues Raised by the Implementation

Implementing the dialogue game has proved to be a very useful way of evaluating the protocol, as it meets the goal of providing a proof of concept by showing that my general theory of persuasion can be conducted via computer mediated dialogues of this form. This implementation has however also raised a number of interesting issues. Below I summarise the three main insights drawn from the evaluation of the implemented dialogue game protocol:

1. The system, acting as referee, cannot use pre-conditions based on mental states of the participants: it infers these from the moves the players make. This means that the pre-conditions to allow a move may be different from those required to sincerely make a move.

2. Natural dialogue is very flexible. Giving support to interactions modelled on natural language utterances requires constraints, and what constraints are appropriate depends on context and purpose. The protocol may impose too few constraints to allow scope for useful computer support.

3. Goodwill and some co-operation is required to make sensible progress and this is again due to the fact that natural language dialogue is so flexible. Thus, unco-operative players can abuse the protocol to stultify the interaction.

These points are also found in other empirical work, such as [172], where often participants were mystified by the effects of the protocol and artefacts of the protocol could be exploited to win the dispute.

Indeed it is this flexibility that presents problems in natural dialogue. Correctly interpreting the force of particular utterances and deciding how best to respond can lead to misunderstandings, arguments at cross purposes, and inefficiencies both in natural dialogue and in its computational representation.

Given these problems, I have identified an alternative method to better support the construction of arguments about action than by modelling natural dialogue. Instead, the insights drawn from a consideration of natural dialogue — the moves that are required and typical patterns of natural dialogues in particular contexts — can be used to provide a tool which instead of attempting to mimic natural dialogue provides a well-defined and productive route through a dialogue capable of addressing a specific situation. In this way the misunderstanding of the justification can be minimised, and the most pertinent attacks can be made. I have taken this approach in the *PARMENIDES* (for Persuasive ARguMENt In DEmocracieS) system. The system is a web-based mediation tool set in the domain of eDemocracy and it is described in detail in Chapter 6, along with another eDemocracy application of my theory which solely involves autonomous BDI agents.

4.5 Summary

In this chapter I have taken the theory of persuasion in practical reasoning articulated in Chapter 3 and transformed this into a dialogue game protocol called the *PARMA Action Persuasion Protocol*. The protocol was defined by a syntax and an axiomatic semantics to specify how persuasive argument over proposed courses of action can be undertaken by two or more participants. This is done by a proponent putting forward a position for the justification of an action and an opponent making a specified attack on elements of the justification. I also provided a discussion of how the proponent can respond to such attacks and how resolution of a conflict is dependant upon the category into which the attack falls. The chapter concluded with a description and discussion of an implementation the PARMA Protocol in the form of a mediation system implemented in the Java programming language. This implemented system is intended as a proof of concept to provide support for computer mediated dialogue tools to be used by human agents and it is intended as a step along the way to a model for persuasive argument in autonomous agents. In the next chapter I will describe how the underlying theory of persuasion over action can be made computational for use in BDI agents.

Chapter 5

Application To BDI Agents

In this chapter I will show how the model for practical reasoning presented in the previous two chapters can be made computational within the framework of an agent based on the Belief-Desire-Intention architecture. This process of practical reasoning in BDI agents is known as ‘the Deliberation Process’ and it is divided into two phases: option generation and filtering. First, Section 5.1 gives a brief discussion of the background of planning in BDI agents, plus informal descriptions of how my account fits with the BDI planning model. Section 5.2 presents the formal definitions I use to describe how BDI agents can generate a set of presumptive arguments for action. Section 5.3 defines the pre-conditions that need to be satisfied in order for an agent to execute each of the attacks on a presumptive argument. Section 5.4 provides an extension of the definitions supplied in Section 5.3, to show how elements of time, uncertainty and degrees of promotion can be captured by the model. This completes the option generation phase of the deliberation process. Section 5.5 describes the filtering phase by showing how all justifiable arguments can be arranged into a Value-Based Argumentation Framework in order to calculate their dialectical status and determine the best intention for the agent to commit to, given its value preferences. Section 5.6 gives a brief example to demonstrate the approach and Section 5.7 concludes with a summary.

5.1 Informal Introduction

The computational setting for the approach is a multi-agent system, in which the agents form intentions based on their beliefs and desires. This is essentially the standard BDI agent model [168], except a small extension is made by associating each desire with a value, the reason why it is desirable. First I describe how an agent can construct a presumptive justification for action, instantiating argument scheme AS1.

A BDI agent has a set of beliefs about the world (used to instantiate R in AS1), and we can therefore expect it to be able to reply “true”, “false” or “unknown” when queried about the status of a proposition. For a well-formed formula of standard propositional logic, if all the propositions it contains are given a truth value by the agent, the formula will evaluate either to true or false for that agent. In such cases we say the agent believes, respectively disbelieves, the formula. If some propositions are unknown by the agent, there are three possibilities. First it may be that all assignments to the unknown propositions will give models for the formula, in which case the agent believes the formula. Second it could be that no assignment which makes the formula true is possible, in which case the agent disbelieves it. Third it may be that some assignments are models and some are not, in which case we say the agent can *assume* the formula to be true. It is important to be able to allow the agents to make assumptions since the knowledge of a particular agent is typically incomplete.

In many typical BDI architectures, e.g., as in [169], the agent also has a library of plans. These plans are designed to achieve some goal which the agent may wish to bring about in the appropriate circumstances, and they represent *all* the states of affairs the agent can attempt to realise. These goals are the *desires* of the agent (corresponding to G in AS1). The process of practical reasoning is designed to select which desires to commit to achieving, namely to form *intentions*. Each plan (corresponding to A in AS1) will have a set of pre-conditions which when satisfied allow it to be performed, and a set of post-conditions which will become true when the plan is carried out and the agent's beliefs about the world are updated. These post-conditions will include the goal for which the plan is undertaken, but there will typically be additional side effects (corresponding to S in AS1). A plan is a sequence of atomic actions: here I do not consider the details of plans and take the execution of the entire plan to be what I have previously termed "performing an action", since deciding which *plan* to perform is the aspect of practical reasoning which I am addressing here. Values (corresponding to V in AS1) are the reasons why the desires are held worth attempting to achieve: certain states of affairs will be held to promote or demote values, perhaps to differing degrees. In addition to the agent's beliefs and desires, I add to this an additional set of *value functions*, one for each value recognised by the agent. The value function takes a desire as argument and returns some assessment representing the degree to which the value is promoted. Positively valued states indicate a degree of promotion of the value represented by the satisfaction of the desire and negatively valued states represent the degree of demotion of the value represented by the satisfaction of the desire. This assessment could be qualitative such as a simple '+' or '–' as envisaged in Chapter 3, but could be made more specific by returning a real number x such that $-1 \leq x \leq 1$. Thus desires include both states of affairs which are desired to be true and states of affairs which are desired to be false. It is the value function that distinguishes them.

Now let us return to AS1. The current circumstances R are a conjunction of propositions which the agent believes, or can assume. The action A is some plan in the plan library of the agent which has pre-conditions which are, or can be assumed to be, satisfied in R. The circumstances S result from the application of the post-conditions of A to R. The goal G is the desire of the agent associated with the plan, and the value V is the value promoted by the realisation of G. These connections can allow us to discover in which ways the agent can, given its beliefs, plans, desires and values, instantiate AS1.

In BDI terms AS1 becomes:

Given the current situation R, there is a plan A which if performed will bring about S, realising G which promotes V.

As well as instantiating AS1 to produce a presumptive argument for executing A, agents can also attack such instantiations, using the critical questions given in Chapter 3. I now describe the conditions under which agents can pose critical questions, for each of the attacks from my theory of persuasion over action. I begin by giving informal descriptions for reasons of clarity and then I will proceed to more formal definitions subsequently.

For each attack I here give the source critical question, a description of when it can be asked and a rendering of the argument it represents.

- **Attack 1a:**

  Source CQ: Are the believed circumstances true? (CQ1).

  Description: The agent can assume, but does not believe R.

  Argument: R may not be true.

- **Attack 1b:**

  Source CQ: Are the believed circumstances true? (CQ1).

  Description: The agent does not believe R and it believes that some other circumstances Q, incompatible with R, are true.

  Argument: R is not true because Q is true.

- **Attack 2a:**

  Source CQ: Assuming the circumstances, does the action have the stated consequences? (CQ2).

  Description: The agent can assume, but does not believe that executing the plan in R will result in S.

  Argument: The action may not have the desired consequences.

- **Attack 2b:**

  Source CQ: Assuming the circumstances, does the action have the stated consequences? (CQ2).

  Description: The agent does not believe that executing the plan in R will result in S and it believes that executing the plan will result in some other state of affairs T, incompatible with S.

  Argument: The action will not satisfy the desire because it will not have the desired consequences.

- **Attack 2c:**

  Source CQ: Assuming the circumstances, does the action have the stated consequences? (CQ2).

  Description: The agent does not believe that executing the plan in R will result in S and it believes that executing the plan will result in some other state of affairs T, incompatible with S, which will not satisfy the desire.

  Argument: The action will not have the desired consequences and it will have alternative consequences which will not satisfy the desire.

- **Attack 2d:**

  Source CQ: Assuming the circumstances, does the action have the stated consequences? (CQ2).

  Description: The agent does not believe that executing the plan in R will result in S and it believes that executing the plan will result in some other state of affairs T, incompatible with S, in which the desire does not promote the value.

  Argument: The action will not have the desired consequences and it will not promote the value because its consequences do not realise a desire promoting the value.

- **Attack 2e:**

  Source CQ: Assuming the circumstances, does the action have the stated consequences? (CQ2).

  Description: The agent does not believe that executing the plan in R will result in S and it believes that executing the plan will result in some other state of affairs T, incompatible with S, in which the desire demotes the value.

Argument: The action will not have the desired consequences and it will demote the value because its consequences realise a desire demoting the value.

- **Attack 2f:**

  Source CQ: Assuming the circumstances, does the action have the stated consequences? (CQ2).

  Description: The agent does not believe that executing the plan in R will result in S and it believes that executing the plan will result in some other state of affairs T, incompatible with S, in which the desire promotes some other value.

  Argument: The action will not have the desired consequences and it will have alternative consequences which will realise a desire promoting some other value.

- **Attack 2g:**

  Source CQ: Assuming the circumstances, does the action have the stated consequences? (CQ2).

  Description: The agent does not believe that executing the plan in R will result in S and it believes that executing the plan will result in some other state of affairs T, incompatible with S, in which the desire demotes some other value.

  Argument: The action will not have the desired consequences and it will have alternative consequences which will realise a desire demoting some other value.

- **Attack 3a:**

  Source CQ: Assuming the circumstances and that the action has the stated consequences, will the action bring about the desired goal? (CQ3).

  Description: The agent can assume, but does not believe the desire is satisfied in S.

  Argument: The plan may not fulfil the desire.

- **Attack 3b:**

  Source CQ: Assuming the circumstances and that the action has the stated consequences, will the action bring about the desired goal? (CQ3).

  Description: The agent does not believe the desire is satisfied in S because it believes that some other desire is satisfied in S.

  Argument: The plan will not fulfil the desire and it will fulfil some other desire.

- **Attack 3c:**

  Source CQ: Assuming the circumstances and that the action has the stated consequences, will the action bring about the desired goal? (CQ3).

  Description: The agent does not believe the desire is satisfied in S because it believes that some other desire is satisfied in S which will promote the value.

  Argument: The plan will not fulfil the desire but it will fulfil some other desire which promotes the value.

- **Attack 3d:**

  Source CQ: Assuming the circumstances and that the action has the stated consequences, will the action bring about the desired goal? (CQ3).

  Description: The agent does not believe the desire is satisfied in S because it believes that some other desire is satisfied in S which will demote the value.

  Argument: The plan will not fulfil the desire and moreover it will fulfil some other desire which demotes the value.

- **Attack 3e:**

  Source CQ: Assuming the circumstances and that the action has the stated consequences, will the action bring about the desired goal? (CQ3).

  Description: The agent does not believe the desire is satisfied in S because it believes that some other desire is satisfied in S which will promote some other value.

  Argument: The plan will not fulfil the desire although it will fulfil some other desire which promotes some other value.

- **Attack 3f:**

  Source CQ: Assuming the circumstances and that the action has the stated consequences, will the action bring about the desired goal? (CQ3).

  Description: The agent does not believe the desire is satisfied in S because it believes that some other desire is satisfied in S which will demote some other value.

  Argument: The plan will not fulfil the desire and moreover it will fulfil some other desire which demotes some other value.

- **Attack 4a:**

  Source CQ: Does the goal realise the value intended? (CQ4).

  Description: The agent can assume, but does not believe the value is promoted by the fulfilment of the desire.

  Argument: The desire may not promote the value.

- **Attack 4b:**

  Source CQ: Does the goal realise the value intended? (CQ4).

  Description: The agent believes the value is demoted by the fulfilment of the desire.

  Argument: The desire demotes the value.

- **Attack 4c:**

  Source CQ: Does the goal realise the value intended? (CQ4).

  Description: The agent does not believe the value is promoted by the fulfilment of the desire and it believes that the desire promotes some other value.

  Argument: The desire does not promote the value, although it promotes some other value.

- **Attack 4d:**

  Source CQ: Does the goal realise the value intended? (CQ4).

  Description: The agent does not believe the value is promoted by the fulfilment of the desire and moreover it believes that the desire demotes some other value.

  Argument: The desire does not promote the value and moreover, it demotes some other value.

- **Attack 5:**

  Source CQ: Are there alternative ways of realising the same consequences? (CQ5).

  Description: There is another plan, the pre-conditions of which can be assumed to be satisfied, and the post-conditions of which will bring about the same state of affairs.

  Argument: There is an alternative way to bring about the same state of affairs.

- **Attack 6:**

  Source CQ: Are there alternative ways of realising the same goal? (CQ6).

  Description: There is another plan, the pre-conditions of which can be assumed to be satisfied, and the post-conditions of which will fulfil the desire.

  Argument: There is an alternative way to satisfy the desire.

- **Attack 7a:**

  Source CQ: Are there alternative ways of promoting the same value? (CQ7).

  Description: There is another plan, the pre-conditions of which can be assumed to be satisfied, and the post-conditions of which will satisfy another desire which will promote the value.

  Argument: There is an alternative way to promote the value.

- **Attack 7b:**

  Source CQ: Are there alternative ways of promoting the same value? (CQ7).

  Description: The pre-conditions of the plan can be assumed to be satisfied, but the post-conditions satisfy another desire which will promote the value.

  Argument: The plan satisfies an alternative desire, which promotes the value.

- **Attack 8:**

  Source CQ: Does the action have a side effect which demotes the value? (CQ8).

  Description: The post-conditions of the plan can be assumed to realise a state of affairs which demote the value.

  Argument: There are side effects of the plan which may demote the value.

- **Attack 9:**

  Source CQ: Does the action have a side effect which demotes some other value? (CQ9).

  Description: The post-conditions of the plan can be assumed to realise a state of affairs which demote some other value.

  Argument: There are side effects of the action which may demote some other value.

- **Attack 10:**

  Source CQ: Does doing the action promote some other value? (CQ10).

  Description: The post-conditions of the plan can be assumed to realise a state of affairs which promotes some other value.

  Argument: There are side effects of the plan which may promote some other value.

- **Attack 11a:**

  Source CQ: Does doing the action preclude some other action which would promote some other value? (CQ11).

  Description: The post-conditions of the plan preclude the pre-conditions for some other plan whose post-conditions would promote some other value.

  Argument: Executing the plan precludes some other plan which would promote some other value.

- **Attack 11b:**

  Source CQ: Does doing the action preclude some other action which would promote some other value? (CQ11).

  Description: The post-conditions of the plan can be assumed to realise the state of affairs but they do not satisfy an alternative desire which would promote some other value.

  Argument: Executing the plan does bring about the state of affairs, but this state of affairs preclude the realisation of another desire which would promote some other value.

- **Attack 11c:**

  Source CQ: Does doing the action preclude some other action which would promote some other value? (CQ11).

  Description: The post-conditions of the plan can be assumed to realise the state of affairs but they do not satisfy any other desire which would promote some other value.

  Argument: Executing the plan will bring about a state of affairs where no other desire promoting some other value can be realised.

- **Attack 12:**

  Source CQ: Are the circumstances as described possible? (CQ12).

  Description: The agent's set of possible states of the world does not contain the given circumstances.

  Argument: The circumstances are not a possible state of affairs.

- **Attack 13:**

  Source CQ: Is it possible to do the action? (CQ13).

  Description: The plan does not exist in the agent's plan library.

  Argument: It is not possible to execute the plan.

- **Attack 14:**

  Source CQ: Are the consequences as described possible? (CQ14).

  Description: The agent's set of possible states of the world does not contain the given state of affairs.

  Argument: The consequences are not a possible state of affairs.

- **Attack 15:**

  Source CQ: Can the desired goal be realised? (CQ15).

  Description: The agent's set of desires does not contain the given desire.

  Argument: The desire is not a legitimate desire.

- **Attack 16:**

  Source CQ: Is the value proposed indeed a legitimate value? (CQ16).

  Description: The agent's set of values does not contain the given value.

  Argument: The value is not a legitimate value.

I now proceed to give a set of definitions to formalise the above descriptions of how BDI agents can generate argument positions and attack these positions, in accordance with my underlying theory.

5.2 Definitions

The definitions that follow describe how a BDI agent can put forward a position regarding the justification of an action, in accordance with my theory. These definitions are not intended to be prescriptive of any internal representation of the agent's beliefs, plans and goals. It is assumed that an agent using the definitions provided below would be equipped with some suitable internal representation for these elements. The formalism presented here is intended as a way in which agents' beliefs can be expressed independently of their internal representation in order to describe how argument scheme AS1, and attacks upon it, can be instantiated.

Additionally, in the definitions that follow no account is given of the notions of time, degrees of promotion of values and agents' certainty regarding propositions. These have initially been omitted to clarify the basis of the model. However, after the presentation of these basic definitions I will give a short proposal to demonstrate how time intervals, certainty factors and degrees of promotion can be incorporated into the model.

Definition 1: *The Beliefs of an Agent.* The beliefs of an agent *j* is a four tuple $<$W$_j$, A$_j$, D$_j$, V$_j$$>$ where,

W$_j$ represents beliefs of agent *j* about the world;
A$_j$ represents beliefs of agent *j* about actions;
D$_j$ represents beliefs about the desires of agent *j*;
V$_j$ represents beliefs about the values of agent *j*;

Definition 2: *Beliefs about the World.* The beliefs of an agent are used to determine which pre-conditions of plans in the agent's plan library[1] are satisfied.

The beliefs about the world of agent *j* are a set of assignments of truth values to propositions about the world.

Let M denote the set of all agents in the system. Let P denote the set of all propositions such that a proposition $p \in$ W$_j$ for some agent $j \in$ M.

Definition 3: *Beliefs about Actions.* The actions available to an agent consist of plans from the agent's plan library which are composed of one or more actions.

[1] I use the term plan library here as used in a system such as the Procedural Reasoning System (PRS)[68]. Alternatively, I note that the agent might derive its plans by reasoning in the specific context.

The beliefs about action of agent *j* are a set of triples $<\alpha$, Pre$_{\alpha j}$, Post$_{\alpha j}>$ where, Pre$_{\alpha j}$ and Post$_{\alpha j}$ are assignments to propositions.

W$_{j\alpha}$ is the state of the world that *j* believes will result from performing α.

Additionally, *j* may *assume* that α can be performed if all elements of Pre$_{\alpha j}$ can be *assumed to be satisfied* with respect to W$_j$.

The set A denotes the set of all actions such that $<\alpha$, Pre$_{\alpha j}$, Post$_{\alpha j}> \in$ A$_j$ for some agent $j \in$ M.

Definition 4: *Desires of an Agent.* The desires of an agent are the post-conditions of plans from the agent's plan library.

The desires of an agent *j* are a set of pairs $<$*d*, Cond$_{dj}>$ such that,

d is a desire and Cond$_{dj}$ is an assignment to a proposition *p*. The interpretation is that *j* believes that the desire *d* is satisfied if Cond$_{dj}$ is satisfied with respect to W$_j$. The notions of satisfaction and assumed satisfaction for Cond$_{dj}$ are the same as that for Pre$_{\alpha j}$.

The set D denotes the set of all desires such that $<$*d*, Cond$_{dj}> \in$ D$_j$ for some agent $j \in$ M.

Definition 5: *Values of an Agent.* The values of an agent are associated with desires and they give the reasons as to why the agent wants to achieve a particular desire.

The values of an agent *j* are a set of triples $<$*v*, *d*, prom$_{vj}>$ such that,

v is a value,
d is a desire,
prom$_{vj}$ is the promotion of the value.

Following the definitions from Section 3.2, it is assumed a value may be:

(i) promoted through fulfilment of a desire, represented by a positive number,
(ii) demoted through fulfilment of a desire, represented by a negative number, or

(iii) neither promoted or demoted (i.e., neutral) through fulfilment of the desire, represented by 0.

The set V denotes the set of all values such that $<$*v*, *d*, prom$_{vj}$$>$ $\in$ V$_j$ for some agent *j* $\in$ M.

Definition 6: *Notions of Satisfaction of Formulae.*

The agent's full set of beliefs is represented as a well-formed formula of standard propositional logic, in which all the propositions are given a truth value by the agent.

Let satA(Formula, W$_j$) be true if Formula can be assumed to be satisfied with respect to W$_j$. A Formula can be assumed to be satisfied if it is not believed to be false.

For example, if agent *j* holds the belief that proposition *p* is true, proposition *q* is false and proposition *r* is unknown, e.g., W$_j$ $=$ $\{$(*p*, T), (*q*, F), (*r*, U)$\}$, then he can *assume* that the formula $\{$*p* & *r*$\}$ is satisfied.

Let satS(Formula, W$_j$) be true if Formula can be satisfied with respect to W$_j$. A Formula is satisfied if it is believed to be true.

For example, if agent *j* holds the same beliefs as above, then he can state that the formula $\{$*p*$\}$ is satisfied, as this is the only proposition known to be true and $\{$*p* & *r*$\}$ cannot be satisfied.

This allows us to distinguish between the mere assumption that pre-conditions hold and the actual knowledge that they hold. This reflects the fact that agents often need to make assumptions because knowledge is typically incomplete.

Now *j* has a presumptive argument for α if:

there is an $<\alpha$, Pre$_{\alpha j}$, Post$_{\alpha j}$$>$ $\in$ A$_j$ such that:
satA(Pre$_{\alpha j}$, W$_j$);
satA(Cond$_{dj}$, W$_{j\alpha}$), and
there is a $<$*v*, *d*, prom$_{vj}$$>$, such that prom$_{vj}$ $>$ 0.

5.3 Pre-conditions for Making an Attack

I now present the pre-conditions which must be satisfied for an attacking agent to question the presumptive argument presented to it in the opposing agent's initial statement of a position. Each attack derives from a critical question as given in Chapter 3 and may have variants, as previously detailed. All attacks are presented in the form of pre-conditions which must be met by the attacking agent *k* in order for it to perform the attack in question on the presumptions present in the instantiation of the argument scheme, given by agent *j*. If all pre-conditions for the performance of the attack are met then agent *k* can make the attack by presenting its argument. For each attack I give the critical question which motivates it, plus any variant, and a natural language definition of the corresponding argument. In this section I will refer to situations which satisfy an agent's desires as 'goals'. This is because the underlying theory is independent of its realisation in the BDI model.

The definitions for the attacks are as follows:

There is an attacking agent *k* $\in$ M such that $<$W$_k$, A$_k$, D$_k$, V$_k$$>$ and agent *k* may attack the position put forward by agent *j* using the set of attacks subject to the following conditions:

Source CQ: Are the believed circumstances true? (CQ1).

Attack 1a: *Pre-conditions for agent k to make an attack:*

satA(Pre$_{\alpha j}$, W$_k$) and,
not satS(Pre$_{\alpha j}$, W$_k$).

Argument: p may not be true.

Attack 1b: *Pre-conditions for agent k to make an attack:*

not satA(Pre$_{\alpha j}$, W$_k$).

Argument: p is not true.

Source CQ: Assuming the circumstances are true, does the action have the stated consequences? (CQ2).

Attack 2a: *Pre-conditions for agent k to make an attack:*

satA(W$_{j\alpha}$, W$_{k\alpha}$) and,
not satS(W$_{j\alpha}$, W$_{k\alpha}$).

Argument: α may not have the desired consequences.

Attack 2b: *Pre-conditions for agent k to make an attack:*

not satA(W$_{j\alpha}$, W$_{k\alpha}$).

Argument: α will not have the desired consequences.

Attack 2c: *Pre-conditions for agent k to make an attack:*

not satA(W$_{j\alpha}$, W$_{k\alpha}$) and,
for no $<$*d*, Cond$_{dk}$$>$ does satA(Cond$_{dk}$, W$_{k\alpha}$) hold.

Argument: α will not have the desired consequences.

Attack 2d: *Pre-conditions for agent k to make an attack:*

not satA(W$_{j\alpha}$, W$_{k\alpha}$) and,
satA(Cond$_{ek}$, W$_{k\alpha}$) and,
$<$*v*, *e*, prom$_{vk}$$>$ and,
prom$_{vk}$ $\leq$ 0.

Argument: α will not have the desired consequences.

Attack 2e: *Pre-conditions for agent k to make an attack:*

not satA(W$_{j\alpha}$, W$_{k\alpha}$) and,
satA(Cond$_{ek}$, W$_{k\alpha}$) and,
$<$*v*, *e*, prom$_{vk}$$>$ and,
prom$_{vk}$ $<$ 0.

Argument: α will not have the desired consequences.

Attack 2f: *Pre-conditions for agent k to make an attack:*

not satA(W$_{j\alpha}$, W$_{k\alpha}$) and,
satA(Cond$_{ek}$, W$_{k\alpha}$) and,

there is a *w*, $w \neq v$ such that <*w*, *e*, prom$_{wk}$> and,
prom$_{wk} > 0$.

Argument: α will not have the desired consequences.

Attack 2g: *Pre-conditions for agent k to make an attack:*
not satA(W$_{j\alpha}$, W$_{k\alpha}$) and,
satA(Cond$_{ek}$, W$_{k\alpha}$) and,
there is a *w*, $w \neq v$ such that <*w*, *e*, prom$_{wk}$> and,
prom$_{wk} < 0$.

Argument: α will not have the desired consequences.

Source CQ: Assuming the circumstances are true and the actions has the stated consequences, will the action bring about the desired goal? (CQ3).

Attack 3a: *Pre-conditions for agent k to make an attack:*
for no < *d*, Cond$_{dk}$> does satA(Cond$_{dk}$, W$_{k\alpha}$) hold.

Argument: the state of affairs resulting from performing α may not realise the desire.

Attack 3b: *Pre-conditions for agent k to make an attack:*
for no < *d*, Cond$_{dk}$> does satA(Cond$_{dk}$, W$_{k\alpha}$) hold and,
for some *e*, $e \neq d$, satA(Cond$_{ek}$, W$_{k\alpha}$).

Argument: the state of affairs resulting from performing α will not realise the desire.

Attack 3c: *Pre-conditions for agent k to make an attack:*
for no < *d*, Cond$_{dk}$> does satA(Cond$_{dk}$, W$_{k\alpha}$) hold and,
for some *e*, $e \neq d$, satA(Cond$_{ek}$, W$_{k\alpha}$) and,
<*v*, *e*, prom$_{vk}$> and,
prom$_{vk} > 0$.

Argument: the state of affairs resulting from performing α will not realise the desire.

Attack 3d: *Pre-conditions for agent k to make an attack:*
for no $<$ *d*, Cond$_{dk}$$>$ does satA(Cond$_{dk}$, W$_{k\alpha}$) hold and,
for some *e*, *e* $\neq$ *d*, satA(Cond$_{ek}$, W$_{k\alpha}$) and,
$<$*v*, *e*, prom$_{vk}$$>$ and,
prom$_{vk}$ < 0.

Argument: the state of affairs resulting from performing α will not realise the desire.

Attack 3e: *Pre-conditions for agent k to make an attack:*
for no $<$ *d*, Cond$_{dk}$$>$ does satA(Cond$_{dk}$, W$_{k\alpha}$) hold and,
for some *e*, *e* $\neq$ *d*, satA(Cond$_{ek}$, W$_{k\alpha}$) and,
there is a *w*, *w* $\neq$ *v* such that $<$*w*, *e*, prom$_{wk}$$>$ and,
prom$_{wk}$ > 0.

Argument: the state of affairs resulting from performing α will not realise the desire.

Attack 3f: *Pre-conditions for agent k to make an attack:*
for no $<$ *d*, Cond$_{dk}$$>$ does satA(Cond$_{dk}$, W$_{k\alpha}$) hold and,
for some *e*, *e* $\neq$ *d*, satA(Cond$_{ek}$, W$_{k\alpha}$) and,
there is a *w*, *w* $\neq$ *v* such that $<$w, e, prom$_{wk}$$>$ and,
prom$_{wk}$ < 0.

Argument: the state of affairs resulting from performing α will not realise the desire.

Source CQ: Does the goal realise the value intended? (CQ4).

Attack 4a: *Pre-conditions for agent k to make an attack:*
$<$*v*, *d*, prom$_{vk}$$>$ and,
prom$_{vk}$ ≤ 0.

Argument: the desire may not promote the value.

Attack 4b: *Pre-conditions for agent k to make an attack:*
$<$*v*, *d*, prom$_{vk}$$>$ and,
prom$_{vk}$ < 0.

Argument: the desire will not promote the value.

Attack 4c: *Pre-conditions for agent k to make an attack:*
$<$*v*, *d*, prom$_{vk}$$>$ and,
prom$_{vk}$ ≤ 0 and,
there is a *w*, *w* $\neq$ *v* such that $<$*w*, *d*, prom$_{wk}$$>$ and,
prom$_{wk}$ > 0.

Argument: the desire will not promote the value.

Attack 4d: *Pre-conditions for agent k to make an attack:*
$<$*v*, *d*, prom$_{vk}$$>$ and,
prom$_{vk}$ ≤ 0 and,
there is a *w*, *w* $\neq$ *v* such that $<$*w*, *d*, prom$_{wk}$$>$ and,
prom$_{wk}$ < 0.

Argument: the desire will not promote the value.

Source CQ: Are there alternative ways of realising the same consequences? (CQ5).

Attack 5: *Pre-conditions for agent k to make an attack:*
satA(Pre$_{\beta k}$, W$_k$) and,
satA(W$_{k\alpha}$, W$_{k\beta}$) and $\beta \neq \alpha$.

Argument: there is an alternative action β which will realise the same consequences.

Source CQ: Are there alternative ways of realising the same goal? (CQ6).

Attack 6: *Pre-conditions for agent k to make an attack:*
satA(Pre$_{\beta k}$, W$_{k}$) and,
satA(Cond$_{dk}$, W$_{k\beta}$) and $\beta \neq \alpha$.

Argument: there is an alternative action β which will realise the same desire.

Source CQ: Are there alternative ways of promoting the same value? (CQ7).

Attack 7a: *Pre-conditions for agent k to make an attack:*
satA(Pre$_{\beta k}$, W$_{k}$) and,
for some *e*, *e* $\neq$ *d*, satA(Cond$_{ek}$, W$_{k\beta}$) and $\beta \neq \alpha$ and,
$<$*v*, *e*, prom$_{vk}$$>$ and,
prom$_{vk}$ $>$ 0.

Argument: there is an alternative action β, leading to satisfaction of an alternative desire, which will promote the value.

Attack 7b: *Pre-conditions for agent k to make an attack:*
satA(Cond$_{ek}$, W$_{k\alpha}$), *e* $\neq$ *d* and,
$<$*v*, *e*, prom$_{vk}$$>$ and,
prom$_{vk}$ $>$ 0.

Argument: α has a side effect which satisfies an alternative desire, which promotes the value.

Source CQ: Does doing the action have a side effect which demotes the value V? (CQ8).

Attack 8: *Pre-conditions for agent k to make an attack:*
satA(Cond$_{ek}$, W$_{k\alpha}$), *e* $\neq$ *d* and,
$<$*v*, *e*, prom$_{vk}$$>$ and,
prom$_{vk}$ $<$ 0.

Argument: α has a side effect which satisfies an alternative desire, which demotes the value.

Source CQ: Does doing the action have a side effect which demotes some other value? (CQ9).

Attack 9: *Pre-conditions for agent k to make an attack:*
satA(Cond$_{ek}$, W$_{k\alpha}$), *e* $\neq$ *d* and,
there is a *w*, *w* $\neq$ *v* such that $<$*w*, *e*, prom$_{wk}$$>$ and,
prom$_{wk}$ < 0.

Argument: α has a side effect which satisfies an alternative desire, which demotes some other value.

Source CQ: Does doing the action promote some other value? (CQ10).

Attack 10: *Pre-conditions for agent k to make an attack:*
satA(Cond$_{ek}$, W$_{k\alpha}$), *e* $\neq$ *d* and,
there is a *w*, *w* $\neq$ *v* such that $<$*w*, *e*, prom$_{wk}$$>$ and,
prom$_{wk}$ > 0.

Argument: α has a side effect which satisfies an alternative desire, which promotes some other value.

Source CQ: Does doing the action preclude some other action which would promote some other value? (CQ11).

Attack 11a: *Pre-conditions for agent k to make an attack:*
satA(Pre$_{\alpha k}$, W$_{k}$) and,
satA(Cond$_{ek}$, W$_{k\beta}$), *e* $\neq$ *d* and,
there is a *w*, *w* $\neq$ *v* such that $<$*w*, *e*, prom$_{wk}$$>$ and,
prom$_{wk}$ > 0 and,
not satA(Pre$_{\alpha k}$, W$_{k\beta}$) and,
not satA(Pre$_{\beta k}$, W$_{k\alpha}$).

Argument: doing α precludes some other action which would promote some other value.

Attack 11b: *Pre-conditions for agent k to make an attack:*
there is a *w*, $w \neq v$ such that $<$*w*, *e*, prom$_{wk}>$ and,
prom$_{wk} > 0$ and,
for no $<$*e*, Cond$_{ek}>$ does satA(Cond$_{ek}$, W$_{k\alpha}$) hold.

Argument: there is some other desire, which promotes some other value, but the desire is not derivable from the state of affairs *s*.

Attack 11c: *Pre-conditions for agent k to make an attack:*
there is no *e*, $<$e, Cond$_{ek}>$ such that
for *w*, $w \neq v$, $<$*w*, *e*, prom$_{wk}>$ and,
prom$_{wk} > 0$ and,
satA(Cond$_{ek}$, W$_{k\alpha}$).

Argument: if there is some other desire, which promotes some other value, then this desire is not derivable from the state of affairs *s*.

Source CQ: Are the circumstances as described possible? (CQ12)

Attack 12: *Pre-conditions for agent k to make an attack:*
for some $p \in$ Pre$_{\alpha j}$, $p \notin$ W$_k$.

Argument: p is a meaningless proposition, according to the attacking agent[2].

Source CQ: Is it possible to do the action? (CQ13)

Attack 13: *Pre-conditions for agent k to make an attack:*
for some $< \alpha$, Pre$_{\alpha j}$, Post$_{\alpha j}>$, $\alpha \notin$ A$_k$.

Argument: α is not a recognised action by the attacking agent.

Source CQ: Are the consequences as described a possible state of affairs? (CQ14)

[2] This attack is intended to convey that the attacking agent's beliefs do not contain the proposition as stated by the opposing agent, and thus the statement is meaningless to the attacking agent.

Attack 14: *Pre-conditions for agent k to make an attack:*

for some $p \in \text{Post}_{\alpha j}$, $p \notin \text{W}_{k\alpha}$.

Argument: *s* contains a meaningless proposition according to the attacking agent.

Source CQ: Can the desired goal be realised? (CQ15)

Attack 15: *Pre-conditions for agent k to make an attack:*

for some $<d, \text{Cond}_{dk}>$, $d \notin \text{D}_k$.

Argument: *d* is not a meaningful desire for the attacking agent.

Source CQ: Is the value proposed indeed a legitimate value? (CQ16)

Attack 16: *Pre-conditions for agent k to make an attack:*

for some $<v, d, \text{prom}_{vk}>$, $v \notin \text{V}_k$.

Argument: *v* is not a recognised value.

5.4 Time, Certainty Factors and Degrees of Promotion

As stated at the start of Section 5.2 the above definitions say nothing about the temporal and certainty aspects of agents' beliefs about the world, or the effects of actions, or the possibility of desires promoting values to different degrees. Here I will give a short proposal to show how such elements can be incorporated into the model given above.

Definition 1a: *The Beliefs of an Agent.*

The beliefs of an agent remain unchanged from those given in Definition 1 in Section 5.2.

Definition 2a: *Beliefs about the World.*

Following Definition 2 from Section 5.2, an agent's beliefs about the world consist of assignments to propositions about the world. Definition 2 is extended to include degrees of certainty in a belief about a proposition, expressed as follows:

Let *p* be a proposition about the world such that *p* $\in$ P. Let *t* be a time such that *t* $\in$ T where T is the set of all time intervals. Let cert$_{pj}$ = -1 $\leq$ cert$_{pj}$ $\leq$ 1 represent agent *j*'s certainty regarding the truth of proposition *p*.

This is interpreted as: agent *j* believes with certainty cert$_{pj}$ that *p* is true at time *t*. If cert $_{pj}$ = -1, agent *j* believes *p* to be definitely false, if cert$_{pj}$ = 1, agent *j* believes *p* to be definitely true, and if cert$_{pj}$ = 0, agent *j* has no opinion as to the truth of *p*.

Following this we can now query an agent to determine the degree of certainty to which the agent subscribes to a particular proposition *p* being true at a particular time *t*.

Definition 3a *Beliefs about Actions.*

Following Definition 3 from Section 5.2, the actions available to an agent consist of plans from the plan library which are composed of one or more actions and each plan has associated with it a set of pre and post conditions for its execution. Again, following Definition 3, the beliefs about action of agent *j* is a set of triples $<\alpha$, Pre$_{\alpha j}$, Post$_{\alpha j}>$ where, Pre$_{\alpha j}$ and Post$_{\alpha j}$ are assignments to propositions.

The pre-conditions for the execution of α must all be met in order for the action to be executable. In order for a pre-condition to be met the agent must believe the proposition to be true. Following this, there is a threshold for belief in the proposition to represent the degree of certainty in the propositions which the agent requires to be met in order to act as though it were true. This threshold may differ for different actions: greater degrees of risk may be tolerated in some cases rather than others. When this threshold is surpassed the particular pre-conditions for the execution of the plan are taken as met. This is expressed as:

α is an action; Pre$_{\alpha j}$ is a set of pairs $<$*p*, threshold$_{pj}>$ and Post$_{\alpha j}$ is a set of pairs $<$*p*, truth$_{pj}>$, -1 $\leq$ threshold$_{pj}$ $\leq$ 1, and -1 $\leq$ truth$_{pj}$ $\leq$ 1.

Pre$_{\alpha j}$ is a set of pre-conditions for α recognised by agent *j*. The interpretation is that *j* believes that α can be performed at *t* if all elements of Pre$_{pj}$ are satisfied with respect to W$_j$ at *t*.

$<$*p*, threshold$_{pj}>$ is satisfied with respect to W$_j$ if $<$*p*, cert$_{pj}$, *t*$>$ and if threshold$_{pj}$ $>$ 0, then cert$_{pj}\geq$ threshold$_{pj}$, else if threshold$_{pj}$ $<$ 0, cert$_{pj}\leq$ threshold$_{pj}$.

Agent *j* believes that if α is performed at *t*, then for all $<$*p*, truth$_{pj}>$ $\in$ Post$_{\alpha j}$, $<$*p*, truth$_{pj}$, *t*+1$>$ will be an element of W$_j$.

This is interpreted to mean that if agent *j*'s pre-conditions for the execution of the plan all pass the threshold for belief at the current time interval, then the post-conditions for that plan will be true at the next time interval and will then form part of the agent's beliefs. This assumes that plans take one interval of time to execute. Given information as to the time taken to carry out a plan, some later time, when the post-conditions hold, could be specified.

Definition 4a: *Desires of an Agent.*

Following Definition 4 from Section 5.2, the desires of an agent *j* are a set of pairs $<$*d*, Cond$_{dj}>$ such that, *d* is a desire and Cond$_{dj}$ is an assignment to a proposition *p*.

The definition of a desire can be extended to include a threshold to be met, regarding the truth of the proposition about the world, in order for the desire to hold for agent *j*. This is expressed as:

d is a desire and Cond$_{dj}$ is a set of pairs $<$*p*, threshold$_{pj}>$. The interpretation is that *j* believes that *d* is satisfied at *t* if Cond$_{dj}$ is satisfied with respect to W$_j$ at *t*.

Definition 5a: *Values of an Agent.*

Following Definition 5 from Section 5.2, the values of an agent *j* are a set of triples $<$v, d, prom$_{vj}>$ such that,

v is a value,
d is a desire,
prom$_{vj}$ is the promotion of the value.

In the statement given previously in Definition 5, the promotion of a value is restricted to three assignments: positive, negative or neutral. This represents respectively, promotion, demotion or neutrality of the value in concern to the particular desire associated with the value. It is possible however, to extend this definition to include a greater degree of precision to which the agent believes that the value is promoted or demoted through fulfilment of the desire. Thus, the definition of a value's promotion of a desire can be expressed as follows:

The values of an agent *j* are a set of triples <v, d, prom$_{vj}$> such that,

v is a value,
d is a desire,
prom$_{vj}$ is a number -1 $\leq$ prom$_{vj}$ $\leq$ 1, representing the degree to which the satisfaction of *d* promotes *v*.

Definition 6a: *Notions of Satisfaction of Formulae.*

Notions of satisfaction and assumed satisfaction of formulae with respect to agent *j*'s beliefs are the same as those given in Definition 6 in Section 5.2.

However, agent *j*'s presumptive argument for performing an action can now be slightly modified. Agent *j* now has a presumptive argument for performing α at time *t* that includes degrees of certainty of propositions, a notion of time and a notion of degree of promotion of values if:

there is an <α, Pre$_{\alpha j}$, Post$_{\alpha j}$> $\in$ A$_j$ such that:
satA(Pre$_{\alpha j}$, W$_j$) at *t*;
satA(Cond$_{dj}$, W$_{j\alpha}$) at *t*+1 and
there is a <*v*, *d*, prom$_{vj}$>, such that prom$_{vj}$ $>$ 0.

I shall now provide a short example to clarify the above definitions. The setting I have chosen for the example is a simple betting card game where the agent is unsure of the outcome of particular actions. Firstly I define the setting and rules of the game.

The particular game I use for this example is the higher/lower card game using one suit only from a standard deck of playing cards. So, the possible cards that can be drawn are 1 - 10, Jack, Queen, King, Ace, with Ace being counted as the lowest card. In this game the dealer reveals a starting card in the first time interval then agent *j* has to place a bet on whether the next card that the dealer reveals in the next time interval will be higher or lower than the previous card dealt. For the purposes of the example I assume that the odds are fixed at 3-to-1 and the agent's actions are restricted to one of two options: betting 10 or not betting 10. The only value that the agent cares about in this particular example concerns its wealth. So, in winning a bet the agent gains 30 and the value 'wealth' is promoted, whereas in losing a bet the agent loses 10 and the value 'wealth' is demoted. We can now instantiate argument scheme AS1 to provide agent *j*'s justification for action in the form of Arg1:

Arg1

R1: Where next card revealed will be higher than current card
A1: betting 10
S1: wins 30
G1: which will increase return
V1: which promotes wealth.

However, this argument is based on the assumption that the next card will be higher than the current one and this outcome cannot always be certain in such a game. So, Arg1 can be attacked on the basis of this assumption. An attacking agent *k* can use attack 1a (assuming that agent *k* can meet the pre-conditions for performing this attack) to state that the circumstances as described may not be true. Here agent *k* is pointing out that agent *j* might not win the bet. In his defence, agent *j* will then have to examine whether his belief that the circumstances hold (corresponding to cert$_{pj}$ from Definition 2a) passes the threshold for assumed satisfaction, as stated in Definition 6a. The threshold could be set according to a number of criteria and in a game such as this it might well be based on the odds of winning. So, the higher the odds are, the lower the threshold for belief would be. In this particular game the odds were given as 3-to-1, so the threshold of belief could be set to 0.33.

Returning to attack 1a on Arg1, if agent *j* can demonstrate that the probability of the next card being higher than the current one exceeds the threshold of 0.33, then he will reject *k*'s attack of 1a and place his bet. Conversely, if the probability of the next card being higher than the current one does not exceed the threshold of 0.33, then agent *j* will accept agent *k*'s attack of 1a and not place a bet. To clarify, consider the following scenarios:

- If the current card is an Ace then there are 12 cards remaining, none of which could possibly be lower than an Ace, and so the probability of a higher card being drawn next is 12 $\div$ 12 = 1. In this case the probability of the next card being higher exceeds the threshold of 0.33 for this particular game and so agent *j* will place a bet on this outcome.

- If the current card is a 5 then there are 12 cards remaining, 8 of which could possibly be higher than a 5 and so the probability of a higher card being drawn next is 8 $\div$ 12 = 0.67. In this case the probability of the next card being higher exceeds the threshold of 0.33 for this particular game and so agent *j* will place a bet on this outcome.

- If the current card is a 9 then there are 12 cards remaining, 4 of which could possibly be higher than a 9 and so the probability of a higher card being drawn next is 4 $\div$ 12 = 0.33. In this case the probability of the next card being higher just meets the threshold of 0.33 for this particular game and so agent *j* will place a bet on this outcome.

- If the current card is a 10 then there are 12 cards remaining, 3 of which could possibly be higher than a 10 and so the probability of a higher card being drawn next is 3 ÷ 12 = 0.25. In this case the probability of the next card being higher does not exceed the threshold of 0.33 for this particular game and so agent *j* will not place a bet on this outcome. Any card higher than a 10 will also produce a probability for the next card being higher that does not exceed the threshold and so agent *j* will not place a bet in such cases.

As we can see from the above examples, the agent cannot be sure of the outcome of the action. So, it makes its decision, as to whether or not to place a bet, based upon its degree of belief as to what the next card will be, given its knowledge of the current card.

In addition to the attack of 1a, an attacking agent could also challenge Arg1 on a different basis. For example, by using attack 2e (assuming that agent *k* can meet the pre-conditions for performing this attack), to state that the action will not result in the expected state of affairs and it will actually result in another state of affairs, in which the desire demotes the value i.e., betting 10 will not win anything but actually lose 10 and demote wealth.

The opposing agent can instantiate argument scheme AS1 using this attack to produce Arg2:

Arg2

R2: Where next card revealed will be lower than current card
A2: betting 10
S2: loses 10
G2: which will decrease return
V2: which demotes wealth.

In this situation the original agent must assess whether or not to place the bet according to the degree of promotion/demotion that the potential outcome has on its value of ‘wealth’. If the agent can afford to lose e.g., he is very wealthy already, then he may value an increase in wealth more than a threatened decrease, and so if the return is positive he will choose to place the bet. However, if the agent cannot afford to lose the stake e.g., the stake consists of all the money that he has, then a decrease in wealth demotes the value more than an increase promotes it, and rationally he should not place the bet. In this situation the loss would be more detrimental to the poorer agent than it would be to the richer agent and this would be reflected in the degree of promotion that they each assign to changes in the value of wealth, in this situation. Thus, different agents may act differently in a given situation according to their assessment of the extent to which the different possible outcomes promote the value concerned.

Additionally, the probabilities of determining whether the next card will be higher or lower than the current card will be affected if the game is played iteratively in a number of rounds. If the game is played by not replacing the cards that have already been dealt then the number of cards left available will decrease by one on each round of the game. In this way, the probabilities of a higher/lower card been drawn next will be dependant upon the cards that have been previously dealt and will differ from those given in the above example. This will in turn affect whether or not the threshold for belief will be met and ultimately the agent's likelihood of placing a bet.

The example given here, which includes degrees of certainty in belief, degrees of promotion of values and a simple temporal aspect, is intended to demonstrate how my account can handle such issues. In the BDI agent examples scenarios of Chapters 6, 7 and 8, I shall not use degrees of certainty and promotion. This is because the particular examples I provide can adequately satisfy their purpose of demonstrating how my account can be used effectively in different domain scenarios to reason about an action to be taken, without the need to use degrees of certainty and promotion. However, as it is often the case that agents are situated in non-deterministic environments with incomplete and uncertain information, I do believe that it is important to note that my account can be extended to deal with such situations.

5.5 Argumentation Frameworks

Now that the procedures by which a BDI agent can instantiate and attack a position regarding action have been set out, an agent now needs some method by which it can choose the best action to commit to from the set that withstood the critical questioning process. In describing this method and throughout the rest of this section I will refer back to the 'the Deliberation Process' of a BDI agent to show how the process of practical reasoning I am presenting is compatible with this process.

As discussed earlier in Section 2.3, the mechanism used by a BDI agent to reason about action is described by Wooldridge in [168] as 'the Deliberation Process'. This process is broken down into two phases: option generation and filtering. During the option generation phase the decision-making agent generates a set of possible alternative actions available for execution, given its beliefs and desires. In the model presented above, corresponding to the generation of options agents can generate a set of presumptive arguments for actions, and the critical questions/attacks which can be used against these arguments. Note that in my model the option generation phase permits contributions from other agents, and as well as other options, challenges can also be made to the possibility of options. Such challenges may raise questions, identify areas of incompleteness and uncertainty of the reasoning agent. These critical questions may themselves give rise to arguments attacking the original argument. The formalism given above allows the option generation phase to be completed and the agent can now move on to the filtering phase. To perform the filtering in this model these arguments are formed into a *Value-Based Argumentation Framework* in the manner of [26], an extension of Dung's framework [55] to accommodate arguments based on values, as described in detail in Section 2.5.3. Recall, whereas in [55] an argument is always defeated by an attacker, unless that attacker can itself be defeated, in [26], attack is distinguished from *defeat for an audience*. The preferred extension thus represents the maximal consistent set of acceptable arguments with respect to the argumentation framework and a given value ordering, which is the maximal consistent position for an audience with that value ordering. In [26] it is shown that the preferred extension for a given value ordering is unique and non-empty, provided it contains no cycles in which every argument relates to the same value. The preferred extension will form the intentions of the agent.

Once the proposing agent's position has been stated, the attacking agent may question this position through the use of the attacks from my model, stemming from the critical questions. However, not all of the critical questions will be applicable to arguments across all domains. For example, CQ7 would be of use in the legal domain (as is discussed in the example of Chapter 8), as it concerns the justification of a past action which is taken as a precedent supporting some future action, but it would not be applicable to many other domains. Therefore, the critical questions need to be analysed to discover which ones apply to the domain in question. Once this list has been determined the agent must then check which specific attacks have their pre-conditions satisfied for making an attack on the opposing agent's position. The attacking agent will then go on to actually state the attacks for which all pre-conditions hold. On completion of this phase the argumentation framework can then be modified to include these attacks on the initial position. The original agent may now pose critical questions against any arguments advanced by the opposing agent, extending the framework further. The process continues until no new arguments can be advanced. The filtering process now commences by calculating the preferred extension of the agent in the light of any value ordering information available. The entire 'Deliberation Process', as described above and in relation to my model, is shown diagrammatically in Figure 5.1.

5.6 Example

To illustrate the above procedure I now present a short example of its application, before going on to look at more extensive examples in the next three chapters[3]. The

[3] Note that in this example and also in the three that follow in Chapters 6, 7, and 8, I use informal descriptions to depict the scenarios. For example, I represent the beliefs of the agents by listing them as natural

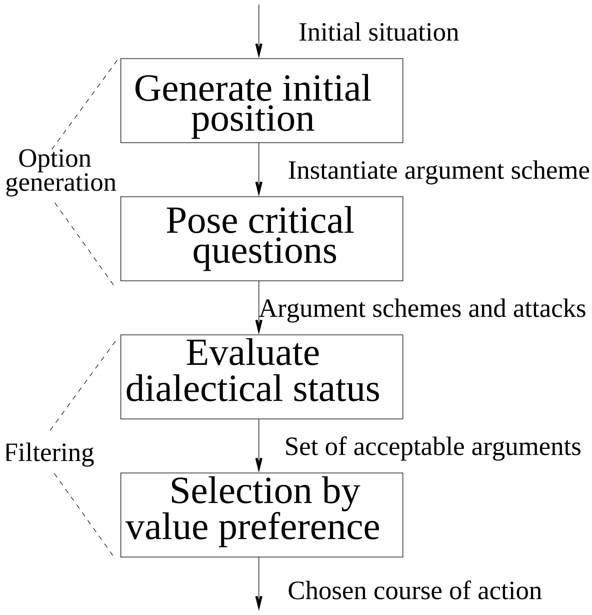

Initial situation
Generate initial position
Option generation
Instantiate argument scheme
Pose critical questions
Argument schemes and attacks
Evaluate dialectical status
Filtering
Set of acceptable arguments
Selection by value preference
Chosen course of action

Figure 5.1: The Deliberation Process.

situation is where two agents are arguing over what action should be taken in a particular context and the dispute is resolved using the processes described above.

The example is taken from a classic moral dilemma discussed by e.g., Coleman [50] and Christie [45]. There are two agents, called Hal and Carla, both of whom are diabetic. However, Hal, through no fault of his own, has lost his supply of insulin and needs to urgently take some to stay alive. Hal is aware that Carla has some insulin kept in her house, but Hal does not have permission to enter Carla's house. The question is whether or not Hal is justified in breaking into Carla's house in order to get some insulin to save his life. The desires are (a) Hal does not die, (b) Carla does not die, and (c) Carla's property (her house and her insulin) remain intact. Both (a) and (b) promote the value 'respect for life' and (c) promotes the value 'respect for property'. Table 5.1 shows the beliefs of the two agents in the initial situation (i.e., before Hal has taken the insulin) and it also shows the beliefs of each individual agent in the consequential situation (i.e., after Hal has taken the insulin). In the table 1 represents "true", -1 represents "false" and 0 represents "unknown", with respect to the agents' beliefs about the propositions.

Hal's view shows the state of the world according to his perspective after he has taken Carla's insulin and Carla's view shows the state of the world according to her

language propositions in tables, rather than using more formal notation and propositional logic representation of the statements. This convention is for ease of presentation and understanding of the examples. However, all arguments and attacks introduced in the examples are in line with the formalism given in Sections 5.2 and 5.3 and if required, they could be represented using the formal notation I have provided.

Table 5.1: **Beliefs of the agents**

Situation	H has insulin	C has insulin	H is alive	C is alive	Property of C intact
Before Hal takes the insulin					
Hal & Carla	-1	1	1	1	1
After Hal takes the insulin					
Hal	1	1	1	1	1
Carla	1	0	0	0	-1

perspective after Hal has taken her insulin. We can see from Table 5.1 that Hal, unlike Carla, knows that he can meet his needs whilst leaving enough insulin for Carla. Hal also intends to replace Carla's insulin, and this is represented by Hal's belief that Carla's property will remain intact.

There are a number of possible desires which may be satisfied by the outcome of the action. These desires either promote or demote the two values in question, which are 'respect for life' and 'respect for property'.

The desires are summarised in Table 5.2. Here 0 represents "unimportant", as the goal is satisfied whether these attributes are true or false.

Table 5.2: **Desires of the agents**

Desires	H is alive	C is alive	Property of C intact	Values: '+' = promoted, '–' = demoted
D1	1	0	0	+life
D2	0	1	0	+life
D3	0	0	1	+property
D4	0	0	-1	–property
D5	-1	0	0	–life
D6	0	-1	0	–life

D1 represents the situation where Hal is alive and thus promotes the value 'respect for life'. Similarly, D2 promotes 'respect for life' as this is the situation where Carla is alive. D3 represents the situation where Carla's property remains undiminished and this promotes the value 'respect for property'. Conversely, D4 is the situation where Hal does break into Carla's house and this demotes the value 'respect for property'. This leaves D5 and D6 and they are respectively the situations where Hal is not alive and Carla is not alive and these both demote the value 'respect for life'.

We can now begin the discussion about the action to be taken and we can assume there are two agents involved in the discussion: Hal and Carla. On the basis of his beliefs Hal can instantiate AS1 to give the following argument, argument Arg1:

Argument Arg1:

R1: Hal has no insulin and will die without some,
A1: so he should break into Carla's house,
S1: to get access to some insulin,
G1: so that Hal remains alive,
V1: promoting the value of respect for life.

Carla's beliefs satisfy the conditions of attack 9, which states that performing this action has a side effect which demotes some other value, this value being 'respect for property'. This instantiates AS1 to argument Arg2:

Argument Arg2:

R2: The insulin belongs to Carla,
A2: so Hal should not break into Carla's house,
S2: so he will not get access to the insulin,
G2: and this will keep Carla's property intact
V2: promoting the value of respect for property rights.

This situation is depicted in the value-based argumentation graph given below in Figure 5.2. In this figure and in all the figures that follow, nodes represent arguments. They are labelled with the given argument identifier, the associated value, and on the right hand side, the agent introducing the argument. Arcs are labelled with the number of the attack they represent.

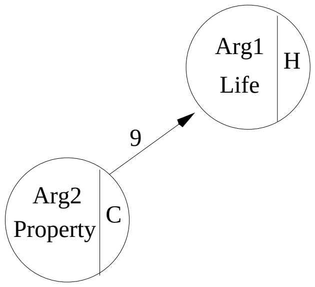


Figure 5.2: Framework 1 of the insulin dilemma.

Carla's beliefs also satisfy the conditions for attack 8, which states that there are unconsidered consequences of the action. This instantiates AS1 with argument Arg3:

Argument Arg3:

R3: The insulin belongs to Carla,
A3: if Hal takes the insulin,
S3: this will leave Carla without any,
G3: which this will threaten her life,
V2: demoting the value of respect for life.

This adds a new node to the graph, as shown below in Figure 5.3:

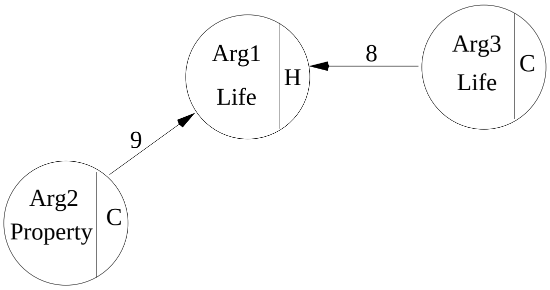


Figure 5.3: Framework 2 of the insulin dilemma.

Carla may herself use attack 2a on argument Arg3, since as Table 5.1 shows, she has only assumed that Hal will not leave her with enough insulin, giving the attacking argument A4:

Attack A4: It is only an assumption that if Hal takes Carla's insulin she will be left with none and die.

From Table 5.1 we can see that Hal knows that Carla has plenty of insulin and so he can make the stronger attack 2b, since the propositions that Carla has insulin and Carla is alive following the action are both true rather than unknown. This gives the attacking argument A5:

Attack A5: It is only an assumption that if Hal takes Carla's insulin she will be left with none and die but we actually know that Carla has ample supplies of insulin and will not die if Hal takes some.

Since both these arguments concern matters of fact they are given the value 'truth', as in [26]. Since 'truth' is always given the highest ranking among values (we can't choose what is the case), this now means that argument Arg3 is defeated, as shown in Figure 5.4:

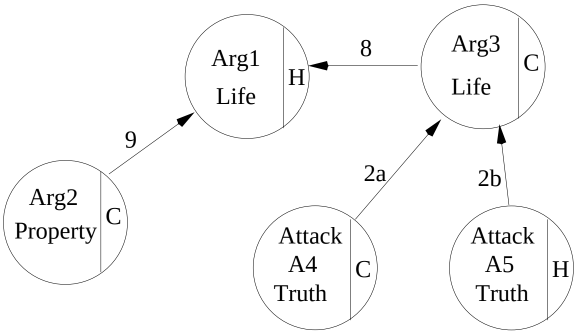


Figure 5.4: Framework 3 of the insulin dilemma.

After attack 2b has been introduced to defeat argument Arg3, we are left with two arguments, Arg1 and Arg2. The conflict between Arg1 and Arg2 is resolved in accordance with the methods described in [26] by calculating the preferred extensions relative to the possible value orders. If we give ‘respect for life’ a higher priority than ‘respect for property’, Hal may take the insulin, whereas if property is respected over life he may not. Thus, the acceptance of Arg1 is subjective and depends on how the audience to the debate ranks the two values involved.

However, [45] proposes that Hal compensate Carla (were he to take her insulin) and this would make use of attack 6 to instantiate AS1 with argument Arg6 as follows:

Argument Arg6:

R6: When Hal has taken Carla’s insulin,
A6: Hal should compensate Carla,
S6: to replenish her supplies,
G6: so that Carla has insulin,
V6: promoting the value of respect for property.

This new argument is added to the framework, as seen in Figure 5.5.

Now, whatever the value ordering, Arg1 and Arg6 are accepted (as when there is an attack on the same value, the attacker always succeeds, as discussed in Section 2.5.3) and so Hal can take the insulin, but he must compensate Carla by replacing the insulin.

This short example is intended to illustrate how the formalism for representing practical argument in BDI agents, presented earlier in this chapter, can be combined with a well understood method for resolving any conflicts over values, through the use of Value-Based Argumentation Frameworks. Thus, this takes forward the model of

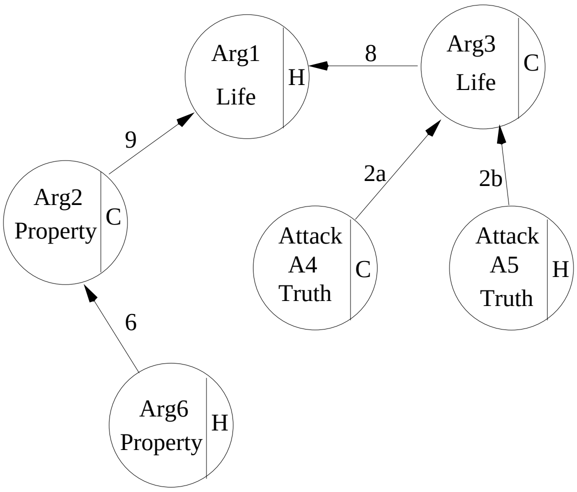


Figure 5.5: Framework 4 of the insulin dilemma.

practical reasoning presented in the preceding chapters of this thesis by providing a qualitative mechanism for generation of agent intentions.

5.7 Summary

In this chapter I have given details of how my model for persuasion over action can be made computational within the framework of agents based on the Belief-Desire-Intention model. I did so by presenting formal definitions to describe how BDI agents can generate sets of presumptive arguments and attacks to form a list of candidate actions to consider. I then described how an agent can choose one of these actions to commit to through the construction and evaluation of competing arguments into a Value-Based Argumentation Framework. I supplemented these definitions with informal descriptions and a short example to illustrate the approach. In the next three chapters I will apply the methods described here to representations of example scenarios in three separate domains where a decision needs to be taken, each involving an action with competing options.

Chapter 6

Application to eDemocracy

In this chapter I present two example applications of my theory of persuasion over action in use in scenarios from the eDemocracy domain. Both applications are illustrated using an example of a recent political debate involving the Government's justification of a proposed action. The first example is an implemented system, realised in the form of an interactive web-based discussion forum. This program implements the theory of persuasion over action in a structured fashion and is intended to address some of the problems identified with the Java program, as discussed in Section 4.4. In Section 6.1 I illustrate this system by first giving some background motivations as to why it is applicable to the eDemocracy domain. I follow this with a description and evaluation of the system, given in Section 6.2.

The second example application applies the definitions and methods described in Chapter 5 to the same specific example debate from the eDemocracy domain. Section 6.3 gives a brief introduction to the application. Section 6.4 presents the BDI agent representation of the debate in terms of the account given in Chapter 5. Section 6.5 discusses some interesting issues that arise from the example regarding the strength of the arguments and how they can be accrued to put forward a more convincing case for justifying the action. Section 6.6 concludes with a summary of both example applications.

6.1 Background

The last two decades have seen a *deliberative turn* in the study of democracy in political philosophy [34]. Prior theories of democracy viewed ordinary citizens as no more than passive consumers of political information and argument, acting only when called upon to vote. In contrast, deliberative theories view citizens as producers of information, engaging as consenting and rational participants in reasoned argument with one another and with their political representatives. As Michelman [114] wrote:

> "Deliberation . . . refers to a certain attitude toward social cooperation, namely, that of openness to persuasion by reasons referring to the claims of others as well as one's own. The deliberative medium is a good faith exchange of views — including participants' reports of their own understanding of their respective vital interests — . . . in which a vote, if any vote is taken, represents a pooling of judgments." [114, p. 293]

Thus, in this view, democracy is not simply a matter of periodic voting: it should also engage its members in informed debate about issues of concern. In a democracy, governments should not only be accountable for the decisions they take, but should justify these decisions in full awareness of, and in response to, the wishes and convictions of the people they govern. Such justification, of course, requires communication between the Government and the people. Today, with the opportunities provided by the World Wide Web, communication is physically easier than ever before, but the long-standing problems that bedevil the effectiveness of communication remain. To be effective, communication must be clear, unambiguous and structured so that misunderstandings are minimised. The structure for persuasive argument which I detailed in Chapter 3 is intended to ease these communication problems, and to promote informed debate. In the next section, I describe a program which exploits this structure, and I illustrate it with an example.

This work complements recent research on the application of information technologies to support democratic participation and debate. Systems such as Zeno [71] and DEMOS [101] aim to assist citizens in communicating with one another and with public officials over matters of community concern and to do so in a dialogue possibly involving multiple simultaneous parties. Likewise, the intelligent systems proposed in [105] for public discussion as part of decision-making on environmental and health regulations are also intended for multi-party dialogues. By contrast, the system implementing the PARMA Protocol, discussed in Section 4.4, is intended for dialogues involving only two simultaneous parties. All these systems seek to embody deliberative notions of democracy and to support public participation in decision-making [72], however, they are not without problems. As discussed in Section 4.4.3, the Java program which implements the PARMA Protocol revealed a number of usability problems: essentially there is too much freedom of expression provided, and hence an overwhelming variety of options for users to select between, and exercising that freedom intelligently requires understanding of the underlying model. These are exactly the problems encountered by earlier systems which have attempted to support democratic debate and dialogue, such as Zeno [71] and TDG [24]. For these reasons, if effective support is to be given to enable the general public to express their views as cogently as possible, some simpler form of interaction is required. The system described here addresses these usability problems by leading the user through a set sequence of moves representing a sensible interaction; by constraining the choice of the user, the need for the user to understand the underlying model so as to make informed selection of moves is removed. Additionally, wherever possible, statements are presented for approval or disapproval, reducing the problems associated with expressing the content of the various locutions. PARMENIDES is intended to realise these objectives.

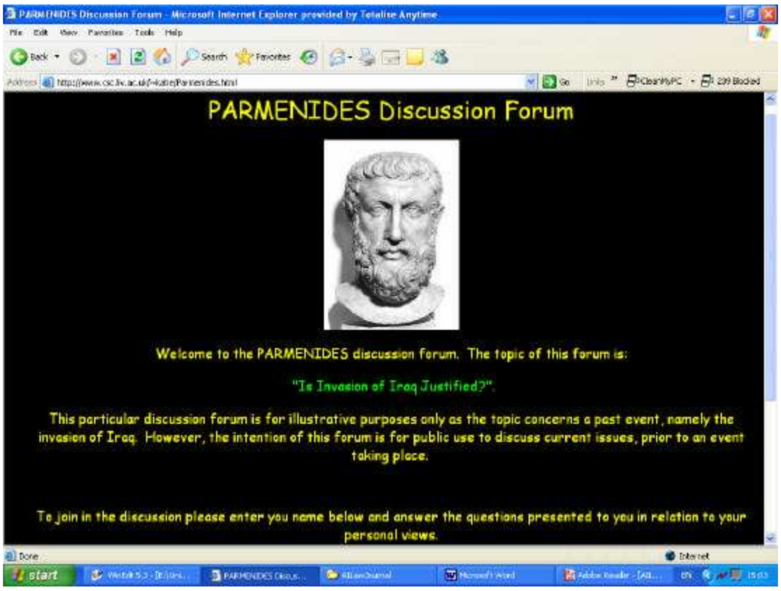
The idea is to provide a simple web based interface which will guide the user in a structured fashion through a justification of an action giving opportunities to disagree at selected points. Each of these disagreements will represent one of the attacks from my theory of persuasion over action, so that the exact nature of the disagreement can be unambiguously identified by the system. The users' responses are written to a database so that information as to which points of the argument are more strongly supported than others can be gathered. Once the original position has been subjected to this critique, another sequence enables users to propose positions of their own, again in a way which will lead them to construct their position in the form of the argument scheme AS1, as detailed in Chapter 3.

In the next section I will describe PARMENIDES, using an example based on the debate from the year 2003 as to whether the UK should go to war with Iraq (set before Iraq was invaded). This was perhaps one of the most widely debated issues of recent years, as well as being one of most controversial debates to feature on the recent international political agenda. Not only did this issue spark debate at national levels, it also received a great amount of time and attention at an international level and disagreement as to the motives and justification of the action taken remains to this day. Debates of such importance require clarity about the issues and any arguments advanced by parties need to withstand critical arguments in order to be justifiable. The aim of the PARMENIDES system is to provide a tool to enable such arguments to be clarified, criticsed and justified.

6.2 The PARMENIDES System

Recall, Section 4.4 described a Java program which was implemented to embody the PARMA Protocol articulated in Chapter 4. The evaluation of the implemented dialogue game revealed a number of undesirable features that emerge through attempting to model natural dialogue. These issues were discussed in Section 4.4.3. Due to the obvious difficulties inherent in trying to computationally represent natural dialogue some of the problems encountered were anticipated before the implementation was embarked upon (such as issues relating to the semantics and relevance of utterances). Conversely, other unanticipated problems with the implementation did also arise. In order to address many of the expected and unexpected problems with the dialogue game described in Section 4.4 I now present a different system, named PARMENIDES, which is also built upon the theory of persuasion in practical reasoning and is embodied by the PARMA Protocol. The specific situation chosen for the setting of this tool is one to facilitate the area of eDemocracy. I describe the system in detail in the following subsections.

6.2.1 Navigation of PARMENIDES


PARMENIDES Discussion Forum - Microsoft Internet Explorer provided by Totalise Anytime
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://www.csc.liv.ac.uk/~katie/Parmenides.html
Go
Links
ClearMyPC
239 Blocked
Done
Internet
start
PARMENIDES Discus...
Microsoft Word

PARMENIDES Discussion Forum

Welcome to the PARMENIDES discussion forum. The topic of this forum is:

"Is Invasion of Iraq Justified?".

This particular discussion forum is for illustrative purposes only as the topic concerns a past event, namely the invasion of Iraq. However, the intention of this forum is for public use to discuss current issues, prior to an event taking place.

To join in the discussion please enter you name below and answer the questions presented to you in relation to your personal views.

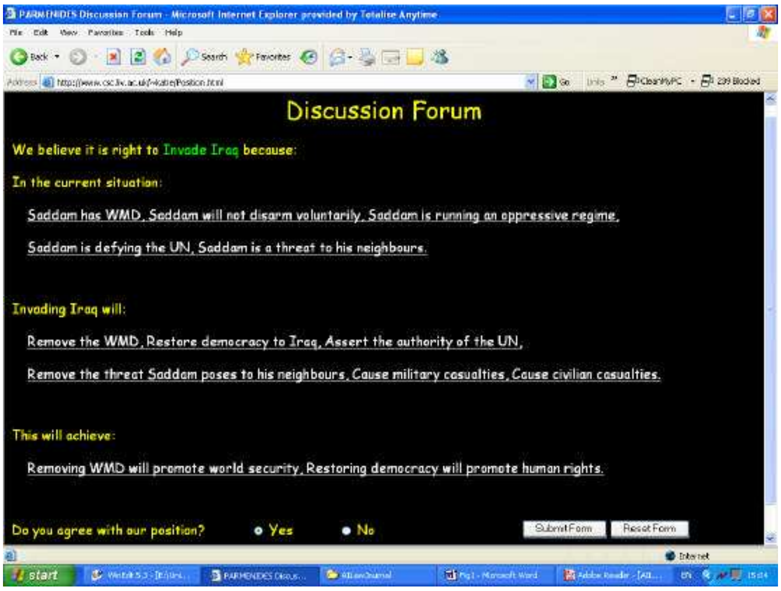
Figure 6.1: Introductory screen.

PARMENIDES is implemented using PHP scripts and provides a highly usable program accessible through any standard web browser[1]. The system can be used at `http://www.csc.liv.ac.uk/∼katie/Parmenides.html`

The aim of PARMENIDES is to present users with a position justifying a particular action and give them the opportunity to critique that position by disputing various points. It does not realise all of the attacks from the theory of persuasion. Some of the attacks that are not used are directed against the soundness of the argument, and here the proponent of the position is relied upon to produce only well formed arguments. Thus attacks 12, 13, 14 and 15 are considered unnecessary, since I assume that the states of affairs and actions described are possible. Additionally, attack 3 is ignored as in presenting the justification for action it is assumed that states of affairs realised by the action will entail the goal (whether the goal is entailed by the consequences should be a matter of logic in this particular scenario), and if the user disagrees with the effects of the action they have the opportunity to pose the similar attack 2. Attacks 6, 7, 9 and

[1] I am most grateful to Sam Atkinson for his invaluable technical help in implementing PARMENIDES.

11 involve the proposal of some counter position, and these are addressed by providing facilities to allow the statement of alternative positions, as will be described in Section 6.2.3. Finally attack 10 is ignored: as discussed previously, this is a subtle matter relating to the motive for an action and since it does not vitiate the proposed action it is not used here.


PARMENIDES Discussion Forum - Microsoft Internet Explorer provided by Totalise Anytime
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.csc.liv.ac.uk/~katie/Position.html
Go Links ClearMyPC 239 Blocked
Internet
start

Discussion Forum

We believe it is right to Invade Iraq because:

In the current situation:

Saddam has WMD, Saddam will not disarm voluntarily, Saddam is running an oppressive regime,

Saddam is defying the UN, Saddam is a threat to his neighbours.

Invading Iraq will:

Remove the WMD, Restore democracy to Iraq, Assert the authority of the UN,

Remove the threat Saddam poses to his neighbours, Cause military casualties, Cause civilian casualties.

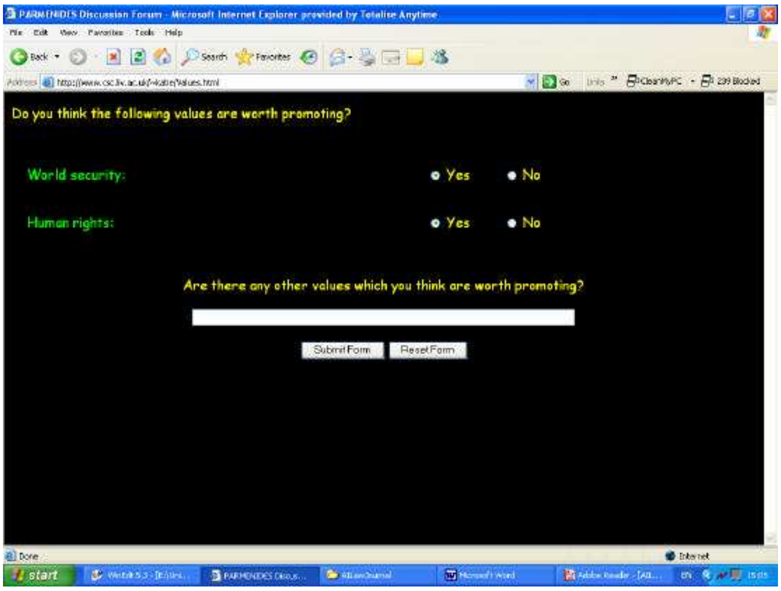
This will achieve:

Removing WMD will promote world security, Restoring democracy will promote human rights.

Do you agree with our position? ○ Yes ● No Submit Form Reset Form

Figure 6.2: Statement of position.

This leaves six attacks to be solicited. This is effected through the navigation of a series of forms. After an introductory screen, Figure 6.1, which takes some information about the user and provides some explanation about the purpose and use of the system, the user is presented with a structured statement of the position to be considered, shown in Figure 6.2. The statement is structured in the form of argument scheme AS1, though the consequences of the action and the goal entailed by these consequences have been combined into one statement for ease of presentation to the user. Because the statement follows AS1 it constrains the provider of the statement to be entirely explicit as to the nature of the argument and the purposes which justify the proposed argument. Note that on this screen, the text which can be agreed or disagreed with is highlighted in white and provides a link to a short justification for the statement. Thus if 'human rights' is clicked, a justification of the UK's commitment to the promotion of human rights will be displayed. At this point users can simply accept the argument, in which case they are sent to a farewell screen. If, however, they wish to challenge the argument, they are sent to a screen concerning values, Figure 6.3, which begins their critique.


PARMENIDES Discussion Forum - Microsoft Internet Explorer provided by Totalise Anytime
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://www.csc.liv.ac.uk/~katie/Values.html
Go
Links
ClearMyPC
239 Blocked
Done
Internet
start

Do you think the following values are worth promoting?

World security: ○ Yes ○ No

Human rights: ○ Yes ○ No

Are there any other values which you think are worth promoting?

Submit Form | Reset Form

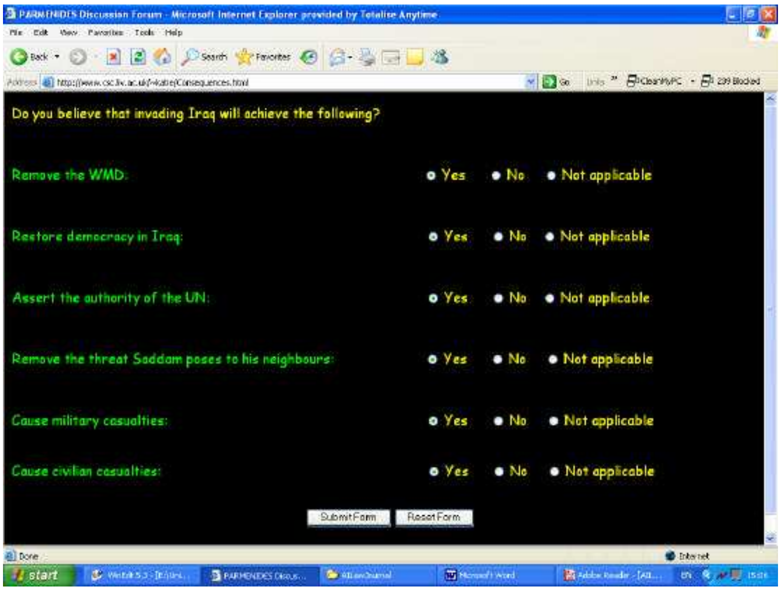
Figure 6.3: Screen presenting social values.

6.2.2 Critiquing the Position

The user can now critique the initial position, starting with the opportunity to make attack 16, registering disagreement with the social purposes underpinning the argument. If the user rejects all such values, further debate is fruitless, since there is insufficient common ground and the exit screen is reached. Assuming that there is at least one value in common, however, he will be taken to a screen which allows him to state whether he believes these values are indeed promoted by the desired consequences of the proposed action (attack 4), with the screen being similar to the one shown subsequently in Figure 6.4. Here he also has the opportunity to state consequences of the action which he believes compromise the desired value (attack 8).

Following this screen the user is invited to agree or disagree that the proposed action will have the consequences envisaged by the proponent. This enables attack 2 and this screen is shown in Figure 6.4. Note that this screen gives the user the chance to check a ‘not applicable’ option box for each statement. This is included to recognise the fact that the user may not be able to agree or disagree about the statements if they are based upon presuppositions about the world that the user does not accept in the first place. For example, the first statement invites the user to say whether he agrees or disagrees that invading Iraq will remove the WMD. However, if he does not believe that Iraq has possession of WMD at all, then he can choose the ‘not applicable’ option. In a later screen the user’s opinions about such presumptions will be elicited, but by including the ‘not applicable’ option here it takes into account that the questions currently being posed may be based upon these as yet unchallenged presuppositions.

Next the user is invited to suggest alternative actions to realise the desired consequences (attack 5), with the screen being similar to Figure 6.3. Finally he is invited to say whether he agrees or disagrees with the description of the current situation (attack 1) and this screen is again similar to that shown in Figure 6.4. The user is then taken to the summary screen which thanks him for using the system and displays the responses that he has given, as partially shown in Figure 6.5.


PARMENIDES Discussion Forum - Microsoft Internet Explorer provided by Totalise Anytime
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.csc.liv.ac.uk/~katie/Consequences.html
Go Links ClearMyPC 239 Blocked
Done
Internet
start
WinEdt 5.3 - [E:/Uni...
PARMENIDES Discus...
AllLawJournal
Microsoft Word
Adobe Reader - [All...
EN
15:06

Do you believe that invading Iraq will achieve the following?

Remove the WMD: ◉ Yes ○ No ○ Not applicable

Restore democracy in Iraq: ◉ Yes ○ No ○ Not applicable

Assert the authority of the UN: ◉ Yes ○ No ○ Not applicable

Remove the threat Saddam poses to his neighbours: ◉ Yes ○ No ○ Not applicable

Cause military casualties: ◉ Yes ○ No ○ Not applicable

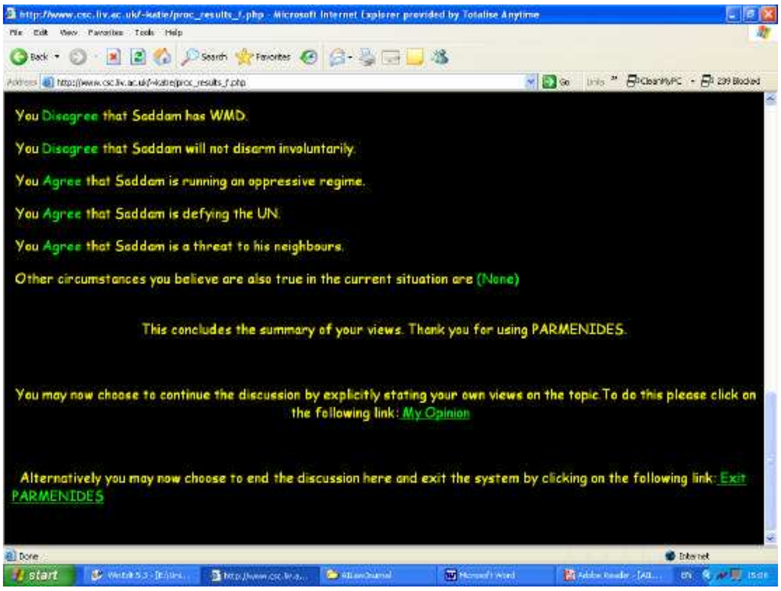
Cause civilian casualties: ◉ Yes ○ No ○ Not applicable

Submit Form Reset Form

Figure 6.4: Links between action and consequences.

6.2.3 Constructing an Alternative Position

Now that the user has supplied all the answers to the questions posed regarding the initial justification for action he was presented with, he is invited to construct his own position regarding the topic in question. However, he may already be satisfied with the answers he supplied when critiquing the original argument and, if this is the case, he may simply choose to exit the system whereupon he will simply be thanked for his input. But, if users do wish to construct their own position on the topic then they are given the opportunity, as shown in Figure 6.5, to follow a link which takes them to a page providing an explanation of the next step. This page explains how their views on the issue will be gathered to construct a justification for an action of the same structure that they were originally presented with upon entering PARMENIDES. Next the user is led to a screen which allows him to enter up to six relevant beliefs



http://www.csc.liv.ac.uk/~katie/proc_results_f.php - Microsoft Internet Explorer provided by Totalise Anytime
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://www.csc.liv.ac.uk/~katie/proc_results_f.php
Go
Links
CleanMyPC
239 Blocked
Done
Internet
start

You Disagree that Saddam has WMD.

You Disagree that Saddam will not disarm involuntarily.

You Agree that Saddam is running an oppressive regime.

You Agree that Saddam is defying the UN.

You Agree that Saddam is a threat to his neighbours.

Other circumstances you believe are also true in the current situation are (None)

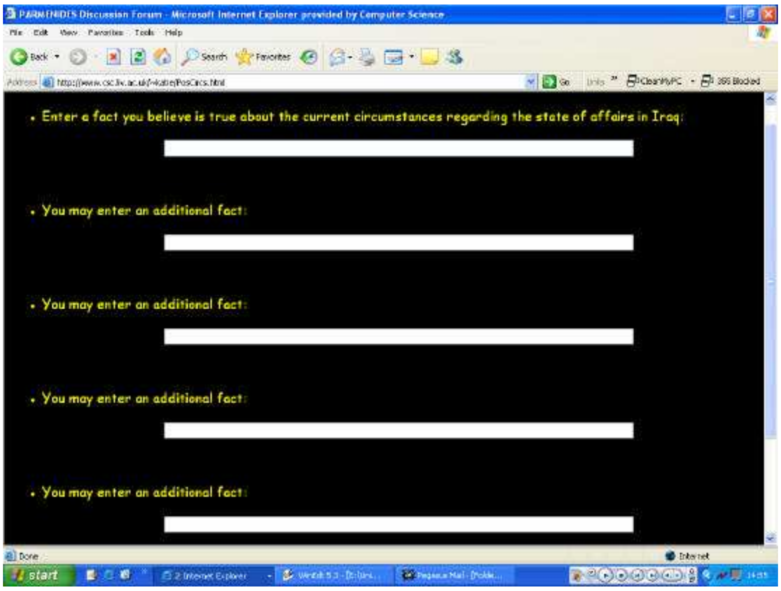
This concludes the summary of your views. Thank you for using PARMENIDES.

You may now choose to continue the discussion by explicitly stating your own views on the topic.To do this please click on the following link: My Opinion

Alternatively you may now choose to end the discussion here and exit the system by clicking on the following link: Exit PARMENIDES

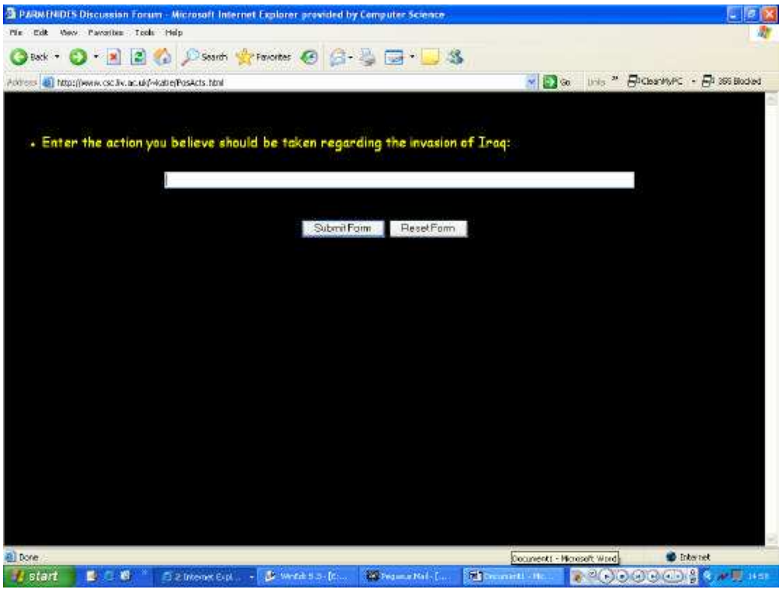
Figure 6.5: Summary screen displaying some of the user's responses and giving him an opportunity to continue or exit the system.

he holds about the current state of affairs in Iraq and this screen is shown in Figure 6.6. Based upon these circumstances, he is then taken to a screen which asks him to state what action he believes should be taken, as shown in Figure 6.7. Following on from this he is asked to input up to six consequences that he believes will follow from executing his specified action, using a screen similar to that shown in Figure 6.6. Finally, another similar screen asks him to enter up to six reasons he believes that the consequences he specified are desirable. All the required questions needed to construct a new position have now been posed so the user is presented with a final screen giving him a summary of the answers he supplied. A partial view of this screen is shown in Figure 6.8. However, it may be the case that the user believes that there are multiple actions which can be executed in the circumstances he specified. If this is the case, then at end of the summary screen he can choose to enter another action, whereby he is taken back to the screen presented in Figure 6.7 and led through the same steps as before until he reaches the summary screen again. This step can be repeated as many times as is required until the user has input all actions, (plus following consequences and reasons) he believes to be desirable in the circumstances he stated. Alternatively, the user may be satisfied that he has submitted his full opinion and, if and when this is case, he can choose to exit the system whereupon he is presented with a final screen. This screen simply thanks him for using the system and then gives him the chance to re-enter PARMENIDES from the start.


PARMENIDES Discussion Forum - Microsoft Internet Explorer provided by Computer Science
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.csc.liv.ac.uk/~katie/PosCircs.html
Go Links ClearMyPC 395 Blocked
Done
Internet
start

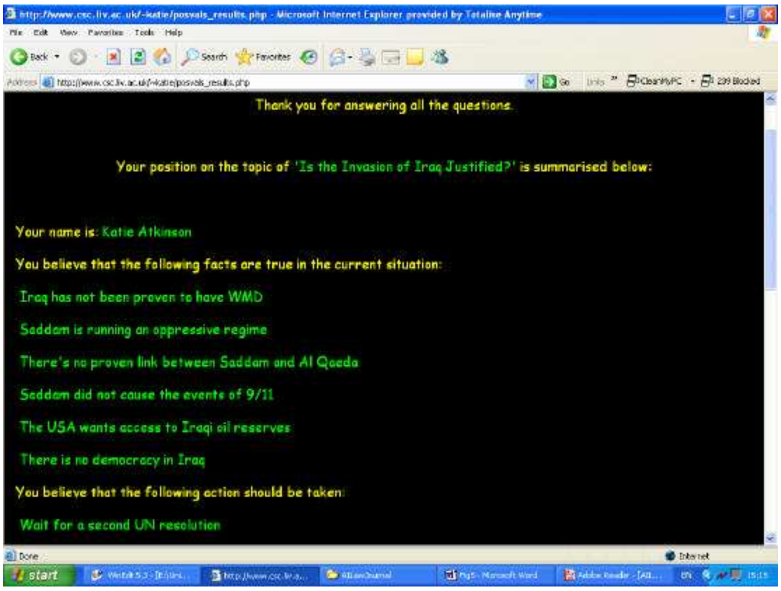
- Enter a fact you believe is true about the current circumstances regarding the state of affairs in Iraq:
- You may enter an additional fact:
- You may enter an additional fact:
- You may enter an additional fact:
- You may enter an additional fact:

Figure 6.6: Screen asking about relevant circumstances.


PARMENIDES Discussion Forum - Microsoft Internet Explorer provided by Computer Science
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.csc.liv.ac.uk/~katie/PosActs.html
Go Links ClearMyPC 395 Blocked
Submit Form
Reset Form
Done
Document1 - Microsoft Word
Internet
start

- Enter the action you believe should be taken regarding the invasion of Iraq:

Figure 6.7: Screen enquiring what action should be taken.


http://www.csc.liv.ac.uk/~katie/posvals_results.php - Microsoft Internet Explorer provided by Totalise Anytime
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.csc.liv.ac.uk/~katie/posvals_results.php
Go
Links
ClearMyPC
239 Blocked
Done
Internet
start

Thank you for answering all the questions.

Your position on the topic of 'Is the Invasion of Iraq Justified?' is summarised below:

Your name is: Katie Atkinson

You believe that the following facts are true in the current situation:

Iraq has not been proven to have WMD

Saddam is running an oppressive regime

There's no proven link between Saddam and Al Qaeda

Saddam did not cause the events of 9/11

The USA wants access to Iraqi oil reserves

There is no democracy in Iraq

You believe that the following action should be taken:

Wait for a second UN resolution

Figure 6.8: Screen giving summary of user's answers constructed into a position stating the justification of his proposed action.

6.2.4 Summary of PARMENIDES

The critique in Section 6.2.2 realises six of the sixteen attacks possible against a position listed in Chapter 3. Each of these attacks proposes no positive information, and thus represents the simplest variant where several variants are possible. Taken together, the six attacks represent a full critique of the position proposed: if none of them can be made, then, provided the position is well formed, the position does indeed represent a justification for the proposed action. Of the ten attacks not provided during this sequence, four challenge the well formedness of the position (which I assume to be in order here), and, apart from the special case of attack 10, which does not dispute the action, the remaining attacks contest the action by developing a justification of an alternative action. The second sequence of screens, allows users to develop such alternative positions, as described in Section 6.2.3.

I am satisfied that PARMENIDES is usable by its target audience, and that it can effectively identify points of disagreement, and record them so that weight of opinion on various issues can be gauged. This is achieved without requiring the user of the system to have any particular familiarity with the underlying model of argument: the attacks are constructed from simple responses without any need for attacks to be explicitly formulated. Using PARMENIDES we can examine the acceptability of various parts of the position. For example, we are able to discriminate between those who support invasion for regime change, from those who are concerned with international security. We can distinguish between those who believe that Saddam has no weapons of mass destruction, from those who believe that he will disarm without invasion, from those who do not believe that he will use the WMD. From this kind of information it is possible for the proponents of the policy to see which elements of the argument need to be put more persuasively or better justified, and which elements could be emphasised to increase the acceptability of the argument.

The free text elements entered by the user are intended to be considered by a moderator who can consider whether they need to be added to the position. Thus if sufficient respondents see some particular circumstance as relevant it can be added to the list of circumstances displayed: if it is not believed by the moderator this is expressed by giving false as its default. Similarly the moderator can examine the proposals for alternatives, and gauge which of these alternatives command substantial support, and the reasons for this.

PARMENIDES was envisaged for use by the Government to assist in the justification of its policy. A similar system could, however, be used by other bodies, such as pressure groups, who could subject their own positions to similar public scrutiny, and solicit additional arguments from the public.

In the next section I provide an example application which again uses the topic addressed in PARMENIDES, but the account given there is a computational application of my theory of persuasion which can be used by BDI agents in accordance with the definitions given in Chapter 5. Subsequent to the presentation of this second application to eDemocracy I will provide a discussion and draw comparisons between the two different representations used to reason about this same topic of debate.

6.3 Application to BDI Agents

In this section I take the computational theory of persuasion over action described in Chapter 5 and apply it to the same particular scenario used in the PARMENIDES system, to show how the theory can be used by autonomous agents. This application models the various participants in the debate as different agents. These agents subscribe to individual beliefs, goals and values, and therefore can represent the different views that can be brought to bear on the problem. I will show the relations between these views and how the arguments can be evaluated through the use of VAFs. However, in this particular example a previously unexplored additional aspect of the theory will be broached and discussed. This is the notion of accrual and how unacceptable arguments can become acceptable for a particular agent through the support of cumulative arguments, which may individually be unacceptable.

6.4 Political Example

The example I will use here involves the same topic that is used in the PARMENIDES system described above, namely: the debate which took place in 2003 as to whether the UK should go to war with Iraq (set before Iraq was invaded). As discussed in Section 6.1, this particular setting has been chosen as it sparked wide debate and argument amongst members of Parliament and the general public. To date, the motivations behind the decision that was taken are still being debated inside and outside of Parliament and thus this seems like a relevant setting to exemplify the use of my model. Again, the example will model the viewpoint of the Government in putting forward its position on the issue and some of the attacks that this justification elicited from members of Parliament and the public.

6.4.1 Context

In the reconstruction of the arguments I will use seven different agents to represent the different views put forward by the parties involved. Firstly there are four agents advocating the action of invading Iraq for different, though sometimes overlapping reasons. These agents will be referred to as: G, representing an agent named George; T, representing an agent named Tony; D, representing an agent named Donald; and C representing an agent named Colin. There will also be three other agents who oppose the action of invading Iraq, again for different reasons. I will refer to these agents as: M, representing an agent named Michael; R, representing an agent named Robin; and J, representing an agent named Jacques.

The application commences by instantiating the agents with the appropriate beliefs, desires and values, in accordance with the definitions given in Chapter 5.

The example makes use of six possible propositions about the world to describe the given situation and these are as follows:

- P1: Saddam has weapons of mass destruction (WMD).
- P2: Saddam is a dictator.
- P3: Saddam will not disarm voluntarily.
- P4: Saddam is a threat to his neighbours.
- P5: Saddam is defying the UN.
- P6: Saddam is running an oppressive regime.

The agents differ quite widely as to which propositions are believed true. Each agent subscribes to the propositions as shown in Table 6.1 with 1 representing belief in

Table 6.1: **Propositions about the world**

Agent	P1	P2	P3	P4	P5	P6
G	1	1	1	1	1	1
T	1	1	1	0	0	1
D	1	0	1	1	0	0
C	0	1	0	0	0	1
M	0	1	-1	-1	-1	1
R	-1	1	0	0	0	1
J	0	1	0	-1	0	1

the proposition, -1 representing disbelief in the proposition and 0 representing unknown to show that the agent has subscribed to neither belief nor disbelief in the proposition.

I now identify the desires and values. We need to identify a set of desires for the agents, and give conditions under which the agents will accept that these desires are realised. We also need to associate these desires with a value, and a degree to which the satisfaction of the desire promotes the value. For now I list the set of desires, conditions and values in Table 6.2 and I will discuss degrees of promotion in the next section.

Table 6.2: **Possible desires and values in the debate**

No.	Desire	Value	Condition to be satisfied
1	No WMD	World Security	Iraq has no WMD
2	No dictator	World Security	Saddam deposed
3	Democracy in Iraq	Human Rights	Saddam deposed
4	International agreement	Good World Relations	All allies agree with the action
5	No human casualties	Respect for Life	No war

Based upon the beliefs and desires given in the above tables, each agent can provide one or more instantiation of AS1. The figures presented subsequently give two argumentation frameworks to show the views of the agents. Initially we can see that there are two values involved in the debate: ‘world security’ and ‘human rights’. The Government’s argument provides two major justifications (which are instantiations of AS1) which endorse the same action of invading Iraq. However, each justification provides different reasons and promotes different values, even though both justifications endorse the same action. We can therefore construct two argumentation frameworks to show the instantiations of AS1 representing the Government’s justifications and the attacks on these instantiations that can be made by the opposing agents. These attacks instantiate the remaining two values of ‘good world relations’ and ‘respect for life’. All these argument schemes, frameworks and attacks are presented below. In the frameworks nodes represent arguments in all the figures. They are labelled with a description of the argument, and on the right hand side, with letters representing the agents introducing the argument. Arcs are labelled with the number of the attack they represent. Following the schemes and frameworks I then summarise what can be deduced from each framework.

6.4.2 Argument Based on Threat to World Security

Firstly I present argument scheme Arg1 showing the Government's first justification of the action, and then the attacks made on it by opposing agents:

Arg1

R1: Saddam has WMD, Saddam is a dictator, Saddam will not disarm voluntarily, Saddam is a threat to his neighbours, Saddam is defying the UN.
A1: we should invade Iraq
S1: which would get rid of the WMD and depose the dictator
G1: so this will remove the threat that Saddam poses to his neighbours and assert the authority of the UN
V1: which will promote world security.

This argument and the attacks that can be made on it by opposing agents given their beliefs and desires are represented in Figure 6.9[2].

Looking at this argumentation framework we can see that the agents subscribe to the following arguments:

Agents G, T and D all put forward Arg1 to justify the action of invading Iraq. The first challenge to be made on this is executed by agent R who uses attack 1a to deny proposition 1 presented in R1 of Arg1. This attack states that agent R does not believe that Saddam has weapons of mass destruction (WMD) and this argument is given the value 'truth', as it is a factual argument. This follows the use of VAFs in [26], where factual arguments are given the value 'truth' and this is ranked as the most important value by all audiences. Also, agent R also does not believe that invading Iraq would get rid of the WMD and so he makes attack 2a as well, which again promotes the value 'truth'. Agent M points out that there is a bad side effect of the action in that the unconsidered value of 'good world relations' will be demoted due to there being international disagreement about the proposed invasion. This is stated in attack 9. This is then attacked by agents T and D who state that they rank the value 'world

[2] As explained in Section 2.5.3, the preferred extension of a VAF is actually computed by removing the unsuccessful attacks from the framework. Figures 6.9 and 6.10 show a diagrammatic representation of all arguments involved in the debate and for explanatory purposes all attacks, including unsuccessful ones, are left in the diagrams, and preferences are shown as arguments.

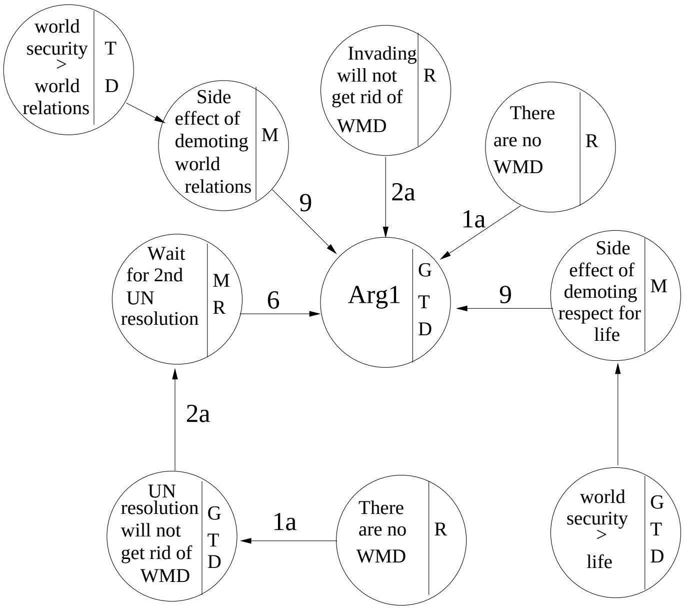

world security > world relations
T
D
Side effect of demoting world relations
M
Invading will not get rid of WMD
R
There are no WMD
R
9
2a
1a
Wait for 2nd UN resolution
M
R
6
Arg1
G
T
D
9
Side effect of demoting respect for life
M
2a
UN resolution will not get rid of WMD
G
T
D
1a
There are no WMD
R
world security > life
G
T
D

Figure 6.9: Arg1 and the attacks on it.

security' higher than they rank the value 'good world relations'. By expressing this preference the value involved in this argument is 'choice' and this is always the least preferred value in the framework [26]. Agents M and R then make a new attack to propose an alternative action to realise the goal. Using attack 6, the alternative action they propose here is waiting for a second UN resolution on the matter. However, this is counter-attacked by all agents supporting Arg1, through attack 2a stating that this newly proposed action will not achieve the goal, as waiting for a second UN resolution will not get rid of the WMD. This argument is then itself counter-attacked by agent R who again uses attack 1a to state that he does not believe that there are any WMD in Iraq in the first place. The final attack on Arg1 is made by agent M who believes that the action will have the detrimental side effect of demoting the value 'respect for life' and he uses attack 9 to state this. However, this is attacked by all proponents of Arg1 through their statement of value preference in which they rank 'world security' as a more important value than 'respect for life', in this situation.

6.4.3 Argument Based on Regime Change

Now that all the agents' arguments have been articulated regarding the justification in Arg1, I now turn to Arg2 in which agents C and T provide a second justification for the same action:

Arg2

R2: Saddam is running an oppressive regime.
A2: we should invade Iraq
S2: to depose Saddam
G2: which will bring democracy to Iraq
V2: which will promote human rights.

This argument and the attacks on it are represented in Figure 6.10:

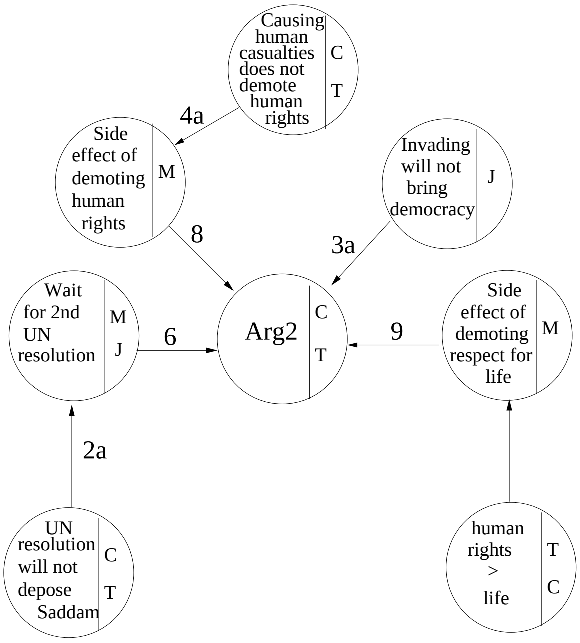

Causing human casualties does not demote human rights
C
T
4a
Side effect of demoting human rights
M
Invading will not bring democracy
J
8
3a
Wait for 2nd UN resolution
M
J
6
Arg2
C
T
9
Side effect of demoting respect for life
M
2a
UN resolution will not depose Saddam
C
T
human rights > life
T
C

Figure 6.10: Arg2 and the attacks on it.

Looking at this above argumentation framework we can see that the agents subscribe to the following arguments:

Firstly, we can see from Arg2 that this argument is based on the belief that Saddam is running an oppressive regime and unlike in the last justification no agent disagrees outright with this fact, as we can see from Table 6.1. So, the first attack made on Arg2 is by agent J who, using attack 3a, states that the action of invading Iraq will not result in a state of affairs that realises the desire of democracy being achieved, since it requires more than just deposing Saddam to achieve democracy. Agent M then makes attack 8 stating that there is a side effect of the action of demoting the value ‘human rights’. This is itself attacked by agents C and T who use attack 4a to state that causing human casualties does not demote the value ‘human rights’. They make this attack as they believe that human casualties may be a necessary evil involved in bringing democracy to a country and the gains more than compensate for the losses, so on balance human rights are promoted. Next, as in the previous framework, attack 6 is used to propose the alternative action of waiting for a second UN resolution and this is put forward by agents M and J. As before, this argument is counter-attacked using attack 2a, this time to state that the alternative action will not achieve the goal of deposing Saddam (as C and T may believe that only military force will remove him) and this is put forward by both agents supporting Arg2. However, this time no agent can attack this argument as agent M did in the previous framework, because they all believe the facts upon which the argument is based to be true. The final attack made on Arg2 is attack 9 in which agent M states that the action of invading Iraq again has the side effect of demoting the value ‘respect for life’. Agents C and T both attack this by stating their belief that the value ‘human rights’ (in relation to the achievement of democracy in Iraq) is more important than the value ‘respect for life’. This concludes the analysis of all the individual attacks used in each framework.

6.4.4 Discussion

It is clear from the above analysis that all agents involved in the discussion have different, but sometimes overlapping reasons for their opinions in the debate. In framework 1 we are able to see that agents G, T and D all accept Arg1 on the basis that they believe Saddam has WMD which he is willing to use to detrimental effect. However, only agents T and D express a value preference of ‘world security’ over ‘good world relations’, which they use to defeat the first instance of attack 9. From this we can see that agent G agrees that there may be the possible side effect of the action of demoting ‘good world relations’, which was pointed out in attack 9. However, he does go on to defend an attack against the second use of attack 9 by stating that he ranks ‘world security’ over ‘respect for life’. From this we can deduce that agent G only needs to have one goal (as opposed to all goals) of Arg1 satisfied in order to justify the action: asserting the authority of the UN is not important to G. As both agents T and D defend all attacks made by the opposing agents, they require all consequences and goals to be satisfied in order for them to be able to justify the action.

Of the opposing agents in framework 1, agent R rejects Arg1 on the basis that he believes the facts upon which it is based are false i.e., there are no WMD. Agent M rejects the argument on a different basis through reasons that reveal he does not support war rather than denying the claim that there are WMD. Agents J and C do not feature in this framework as their views solely relate to arguments about the conveyance of democracy.

In framework 2 we can see that agent T supports this argument in addition to Arg1 and so he is the only agent who sees the need to justify both arguments in order to be able to justify the action. However, agent C also supports Arg2 and as he did not support Arg1 we conclude that he believes Arg2 to be sufficient on its own for the justification of the action.

Of the opposing agents M again reveals his anti-war attitude through the attacks he makes. Agent J disagrees with the result of the action showing that his attitude reflects the belief that democracy will not be achieved through invasion, which is the main thesis of his position. Agent R does not feature in this scenario as he is only interested in arguments resting on the basis of the evidence of WMD.

6.5 Accrual and Strength of Argument

In addition to the individual attacks in the frameworks there is also an attack that can be made between the two argument scheme instantiations Arg1 and Arg2, and this is attack 10. This is shown below in Figure 6.11.

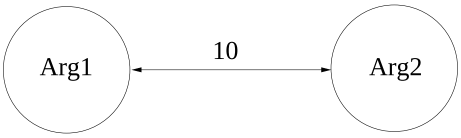


Figure 6.11: Arg1 and Arg2 mutually attack each other.

An agent making attack 10 does not dispute that the action should be performed, but disputes the motive for performing it. In the example, G and D accept Arg1 but not Arg2 and C accepts Arg2 but not Arg1. Thus G may attack Arg2 by saying that regime change is not a justification for invasion, but removing WMD is, and C may attack Arg1 by arguing the contrary. The case of T is different, since he accepts both arguments. If T accepts that both Arg1 and Arg2 are sufficient to justify invasion, he could be challenged to choose between ‘world security’, the value promoted by Arg1, and ‘human rights’, the value promoted by Arg2, so as to clarify his “real” reason for advocating invasion. In practice some politicians seemed to be in the position of T, and generally made the removal of WMD their lead justification, although subsequent to the failure to discover WMD, they cite regime change as sufficient in itself. If, however, desires can promote values more or less strongly, it may be that only one of the arguments is sufficient to justify the action. This will then be the “real” reason, and the other argument is superfluous. A different case is where neither argument is sufficient by itself to promote the action. Here attack 10 is inappropriate, since the two arguments are now intended to be mutually supporting and the action is justified only if both arguments stand. An additional pre-condition for attack 10 is thus that the attacking argument be sufficient to justify the action. The need in some cases to have mutually supporting arguments introduces the notion of accrual. To explore this notion there needs to be some mechanism for distinguishing degrees of promotion, and determining when an argument is sufficient to justify the action. The definitions in Section 5.4 allowed for both these ideas, and I illustrate them with an example in the next subsection.

6.5.1 Degrees of Promotion

The definitions in Section 5.4 represent degrees of promotion as numbers ranging from -1 to 1. This is intended to be flexible enough to accommodate a variety of concrete treatments. Assigning and combining things such as degrees of promotion presents conceptual and practical problems. What I propose here is admittedly rudimentary: I make no claims for its cognitive validity, but use it only to illustrate some points about accrual. It is assumed that agents can make a subjective assessment of the degree of promotion by responding to a question such as “from the standpoint of human rights, how important is it that a country be ruled democratically” with a qualitative assessment such as “utterly, very, somewhat, a little, or not at all”. It is further assumed that satisfying several desires promoting a value will promote that value to a degree greater than is achieved by satisfying only one of them. The treatment is consistent with these assumptions, but remains rather *ad hoc*. The qualitative assessments are translated into numbers as follows:

- if the agent replies “not at all” the value is *not* promoted and we assign the number 0;
- if the agent replies “a little” the value is *weakly* promoted and we assign 0.3;
- if the agent replies “somewhat” the value is *moderately* promoted and we assign 0.5;
- if the agent replies “very” the value is *strongly* promoted and we assign 0.7;
- if the agent replies “utterly” the value is *fully* promoted and we assign 1.

For combinations of two desires promoting the same value, or two arguments proposing the same action the relevant numbers are added then their product is subtracted: i.e., *combine(a,b)=(a+b)–(a*b)*. This is the formula used for rule combination in MYCIN [41]. Again I make no claims for this other than that it satisfies the desired property of increasing the degree of satisfaction while remaining in range. I also, again quite arbitrarily, take 0.7 as the threshold (i.e., the value must be strongly promoted) which must be attained if the action is to be justified, that is if the action is worth performing for the sake of this particular value.

I now return to the example. Agent T needed to accept both Arg1 and Arg2 to convince him to act. Suppose he sees both arguments as moderately supporting the action, both with value 0.5. Separately neither is a sufficient reason to act, but together they support the action at 0.75, and so exceed the threshold. Note that if one argument had been no more than weakly supportive, the combined value would be only 0.65: some third argument offering at least weak support would be required to reach the threshold. In contrast, agents G and D see Arg1 as sufficient to justify the action on its own, and agent C sees Arg2 as sufficient in itself. For these agents there is a single reason to act and attack 10 will lead them to reject the other argument.

This illustrates one form of accrual, where two distinct arguments are involved. Arg1 illustrates a different form of accrual in that it is based on the satisfaction of two desires promoting the same value. If we separate these into Arg1a, based on the removal of Saddam's threat, and Arg1b, based on asserting the authority of the UN, we can see that there are various possibilities. It may be that for some agents one of the desires promotes the value sufficiently to support the action on its own. Suppose that, as for agent G in Figure 6.9, the removal of the threat strongly promotes 'world security'. Then Arg1a provides sufficient justification and G need not defend attacks directed at the part of the goal representing Arg1b. Agents D and T, in contrast, may see both parts of the goal as necessary (because both only moderately promote the value) and so must defend it against all attacks. Thus we can use two distinct types of accrual depending on whether one or more values are promoted. Different values require two arguments and we can ask which provides the *real* motive to act. One value can be represented as a single instance of the argument scheme. In this case, because only one value is involved, an agent who believes that an argument based on one of the desires is sufficient need not reject the argument based on the other desire: the second argument is at worst superfluous and may strengthen the justification for the action.

It is often the case that in everyday practical reasoning we may be unsure as to whether to execute an action or not. However, if we find a separate reason (even if it is weak itself) to execute the action then this can compel us enough to carry out the action. This notion of accrual is also mentioned by Walton in [164] in his argument scheme for the 'Argument From Sign'. Here he states that the more signs that are brought into the scheme, the more inferences can be drawn from them and the more the case builds up. This is particularly relevant for the political forum as the more evidence that is presented to justify an action, the more likely ministers and the public are to accept the argument, enabling the Government to win their support.

Also note that in the definitions given in Chapter 5 degrees of certainty for the individual elements of a position are included to allow representation of varying degrees of belief and value promotion. For example, an agent will only believe some proposition about the world if it passes a certain threshold for belief. Likewise, a value promotion will only be acceptable to an agent if his belief that the desire actually promotes the value passes the set threshold. Again, I have not explored degrees of certainty of individual elements of a position as practical rather than theoretical reasoning is the main focus of the model, but it is worth noting that my model can accommodate this concept. This is something which would be particularly applicable to a domain that involves elements of risk, such as the medical one, as discussed in Chapter 7. For example, it may be crucial to have a strong belief that a treatment given to a particular patient will have the desired effect. Or it may be acceptable to give a patient a treatment with side effects, as long as we are sufficiently satisfied that these side effects will not be too detrimental for the patient's health.

6.6 Summary

In this chapter I have presented two example applications which make use of my theory of persuasion over action within the eDemocracy domain. The first example is a web-based mediation system to solicit public opinion on the Government's proposed justification of an action. This implementation was brought about by a desire to address some of the problems that were presented by the Java implementation of the PARMA Protocol, as discussed in Section 4.4 and PARMENIDES has indeed overcome many of the problems.

The second application is an example to demonstrate how BDI agents can reason about the same topic as that dealt with by the PARMENIDES system, in accordance with the definitions given in Chapter 5. This second account has demonstrated how superficial agreement may conceal subtle but important differences in beliefs and aims.

The general theory of persuasion over action was applied to the political domain to give concrete evidence of how real life issues can be debated with computational agents using this model. An additional point from the BDI agent application was made by demonstrating how such argumentation could make use of the notion of accrual of arguments. This concept is an important feature of persuasive debate and I believe that my model is able to give some insight into this phenomenon. A further discussion of this feature and a comparison between the BDI method of argumentation and the PARMENIDES system will be given in Chapter 9.

Chapter 7

Application to Medicine

In this chapter I apply the definitions and methods described in Chapter 5 to a second example, this time from the medical domain, to show how BDI agents can use my model to reason about medical information. Section 7.1 gives a brief overview of the application. Section 7.2 discusses the domain and motivation for the example. Section 7.3 presents the actual example, in terms of the account described in Chapter 5. Section 7.4 concludes with a summary. The example given here is intended solely to illustrate the method I am advocating: no claims are made for the depth of the medical knowledge.

7.1 Deliberative Reasoning About Medical Treatment

The example application presented here is a system for reasoning about the treatment of a patient using a single BDI agent. The purpose of the application is to show how the incorporation of an argumentation component in this domain can add value to a collection of existing information agents. It is assumed that a number of information sources, representing different areas of medical knowledge and facts about individuals, and different policies and perspectives relevant to the problem are available. The focus of this application is the *Drama* (for Deliberative Reasoning with ArguMents about Actions) agent which orchestrates these contributions in argumentation terms, and comes to a decision based on an evaluation of the competing arguments. First I will discuss the particular application which is used to exemplify the approach[1].

[1] I am most grateful to Sanjay Modgil for his permission to reproduce the domain knowledge for the example given in this chapter, which he contributed in joint work with myself and Trevor Bench-Capon that can be found in [20].

7.2 Background

Clinical guidelines promote best practices in clinical medicine by specifying the selection and sequencing of medical actions for achievement of medical goals. There is a large body of research into computational support for authoring and enactment of clinical guidelines [159]. Authoring tools support specification of a guideline in some suitable knowledge representation formalism. This specification can then be executed in a specific clinical context so as to enforce compliance with the best practice encoded in the guideline. The authored guidelines need to be specified at a level of abstraction that enables enactment in any number of contexts. It is at execution time that the context dependent choice of specific medical actions must be made.

For example, a guideline may indicate that treatment of a patient recovering from myocardial infarct (heart attack) requires realisation of the treatment goals: treat pain; treat sickness; prevent blood clotting. It is at execution time that the specific context must be accounted for, in order to decide which precise action should be chosen for realising each of these goals. Examples of contextual factors that influence the decision include:

- information about the specific patient being treated, e.g., administration of a particular drug for preventing blood clotting may for safety reasons be contraindicated by a patient's clinical history,
- concomitant treatments, e.g., the efficacy of a drug for preventing blood clotting may be reduced by drugs being administered for a gastrointestinal condition,
- local resource constraints, e.g., budget constraints at the local hospital may indicate a preference for one drug over another,
- local organisational policies, e.g., the local health authority may have evidence-based preferences for one drug over another.

Through the application of deliberative argumentation, as I have described in the proposals in this thesis, it is possible to model how contextual factors of the above type can be brought to bear on what is the most appropriate treatment action in a given situation. In particular, by structuring a recommendation for action as an argument instantiating argument scheme AS1, we can effectively account for the influence of contextual factors on the decision making process; i.e., in terms of arguments instantiating AS1's critical questions. Furthermore, the complexity and diverse nature of the contextual knowledge and reasoning suggests distribution and specialisation of knowledge and reasoning resources in medical multi-agent systems. Each resource represents a source of arguments and brings its own perspectives, goals and values to the decision making process.

In the example below, the medical knowledge cited is for illustrative purposes only: I make no claims for it either as a model of the medical domain, or as a representation of the state of the art of medical systems. The purpose here is only to show how value can be added to this domain by the addition of an argumentation agent capable of reasoning with multiple perspectives and drawing on a range of sources.

7.3 Example Application

In this application the Drama agent will be capable of operating with standard information agents. Therefore all argumentation knowledge will be located inside the Drama agent, and the other agents can be regarded as functioning as conventional knowledge and database systems. In particular these other agents need have no knowledge of values.

The other agents that the Drama agent will interact with in the example are shown in Table 7.1. Some will contain generic medical knowledge, while others are specific to the organisation. All can be quite limited in scope: it is the Drama agent that will supply the bigger picture, bringing the contributions together and organising and evaluating them. If desired, however, these other agents could be more sophisticated: for example the Cost Agent could negotiate with suppliers to price the drugs, and perhaps also negotiate a budget for treatment instead of being a static repository of prices and budget information. The Policy Agent could use argumentation and external information in the same way as the Drama agent. If these more sophisticated resources were available, the Drama agent would be at the heart of a true multi-agent system. Since, however, the wish is to concentrate on the Drama agent itself, and as it need make no assumptions about the other components, I take them here to have their simplest form.

Table 7.1: **Agents in the Drama System**

Agent	Type	Scope
Treatment Agent	Knowledge base	Generic medical policy and knowledge
Policy Agent	Knowledge base	Organisation specific knowledge
Safety Agent	Knowledge base	Generic medical knowledge
Patient Agent	Database	Patient specific information
Cost Agent	Knowledge base	Organisation specific knowledge
Efficacy Agent	Knowledge base	Specific medical knowledge

As in the general approach described in Chapter 3, the Drama agent will use critical questions (pertinent to this particular example) to generate arguments. Note that unlike in the example in the previous chapter, here I discuss the questioning of the presumptive arguments in terms of critical questions, as opposed to in terms of attacks. This is due to the nature of this particular example where the reasoning involved is of a more deliberative nature effected internally inside a single agent, the Drama agent. In dialogical interactions it is more natural to think of criticisms directed against an opponent in terms of attacks. Conversely, in monological reasoning, such as will be carried out by the Drama agent, it seems more natural to describe the process as critical questioning where the agent is attempting to convince itself of the best action to take, given its beliefs. Although I will describe the process here in terms of critical questioning, all the conditions that are needed to satisfy a criticism of a position, as described in Chapter 5, are still applicable. As all critical questions relate to attacks from these definitions, any critical questions used in this example must satisfy the minimum conditions for posing the criticism, as given in Chapter 5.

Again, as in the previous example, I take all agents to have a common representation, and any knowledge claimed by the agents to be true. Given these assumptions there are six critical questions pertinent to this particular application:

- CQ1: Are the believed circumstances true?
- CQ3: Will the action bring about the desired goal?
- CQ5: Are there alternative ways of realising the same consequences?
- CQ6: Are there alternative ways of realising the same goal?
- CQ8: Does doing the action have a side effect which demotes the value?
- CQ9: Does the action have a side effect which demotes some other value?

The Drama agent now constructs an argumentation framework by instantiating AS1 and posing these critical questions. I will illustrate the operation of the system with a running example of a patient whose health is threatened by blood clotting. The framework begins with the null option - do nothing (EA0) as shown in the instantiation of AS1 below.

EA0 R0: Where the patient is likely to have blood clotting
A0: we should do nothing
S0: which means the patient will have blood clotting
G0: so blood clotting will not have been prevented
V0: and this will not promote efficacy.

The purpose of EA0 is similar to the assumption of the negation of the desired goal in refutation resolution: extensions of the resulting argument frameworks will be acceptable only if they do not contain this argument. The goal of preventing blood clotting is now passed to the *Treatment Agent*.

The Treatment Agent is one among a number of treatment agents, each of which is specialised for recommending treatment actions for a medical speciality. In this example, the Treatment Agent is specialised to reason about the cardiac domain. This Treatment Agent could be at any level of sophistication, provided that it has knowledge of medical actions (e.g., drug administrations) and their effects, and the clinical goals that these effects realise, i.e., knowledge of the type required to instantiate argument scheme AS1. This would require access to guideline knowledge indicating the clinical goals to be realised (and the scheduling of these goals), as well as more detailed medical causal knowledge (e.g., drugs and their effects) of the type encoded in remotely accessible medical terminologies of the type described in [139]. Let us assume the required knowledge is encoded locally in the Treatment Agent as a Prolog knowledge base. This knowledge base might include (in the following, the variable X stands for the patient to whom the clinical reasoning is being applied):

```
prevent_blood_clotting(X):-
  reduce_platelet_adhesion(X).

prevent_blood_clotting(X):-
  increase_blood_clot_dispersal_agents(X).

reduce_platelet_adhesion(X):-
  not contraindicated(aspirin,X),
  prescribe(aspirin,X).

reduce_platelet_adhesion(X):-
  not contraindicated(chlopidogrel,X),
  prescribe(chlopidogrel,X).

increase_blood_clot_dispersal_agents(X):-
  not contraindicated(streptokinase,X),
  prescribe(streptokinase,X).
```

The Treatment Agent will therefore be able to return the information that blood clotting can be prevented by reducing platelet adhesion, which can, assuming aspirin is not contraindicated, be achieved by prescribing aspirin. The Drama agent can use this information to instantiate AS1, thus providing a justification for this action, i.e., that it will reduce platelet adhesion, which realises the goal of preventing blood clotting, and so is an efficacious plan.

EA1 R1: Assuming no contraindications
A1: we should prescribe aspirin
S1: which will reduce platelet adhesion

G1: preventing blood clotting
V1: and so promotes the value of efficacy.

This argument has to be subjected to a critique to ensure that there are no better alternatives. The Drama agent will go through its repertoire of critical questions. Posing CQ5 will ask for alternative solutions to reduce platelet adhesion from the Treatment Agent and elicit the information that chlopidogrel will also reduce platelet adhesion. Asking CQ6 will seek further solutions from the Treatment Agent for preventing blood clotting and will identify the alternative course of action of administering streptokinase, which has the same goal of preventing blood clotting, but via a different effect of increasing the blood's production of agents that disperse clots. These are formed into two arguments, EA2 and EA3:

EA2 R2: Assuming no contraindications
A2: we should prescribe chlopidogrel
S2: which will reduce platelet adhesion
G2: preventing blood clotting
V2: and so promotes the value of efficacy.

EA3 R3: Assuming no contraindications
A3: we should prescribe streptokinase
S3: which will increase blood clot dispersal agents
G3: preventing blood clotting
V3: and so promotes the value of efficacy.

These three arguments all mutually attack one another and they also all attack EA0, using CQ8, to state that the action of EA0 has effects which demote the value of 'efficacious action'. These arguments all give rise to the argumentation framework shown in Figure 7.1. Note, unlike in the previous chapter's example all the reasoning involving the action to be taken is effected inside a single agent. Thus, in these frameworks nodes are labelled with only the instantiation of AS1 that they represent and the value involved in that justification. Arcs are labelled with the critical question that is challenging the argument.

Any of EA1, EA2 or EA3 would serve to defeat EA0, 'do nothing'. However, they are in mutual conflict. As they all relate to the same value (and the preferred extension is empty for all audiences), there is a free choice between them. They can be chosen according to intrinsic preferences regarding the goal or the actions themselves. The Drama agent therefore contacts the *Policy Agent* to see what the preferences of the organisation are.

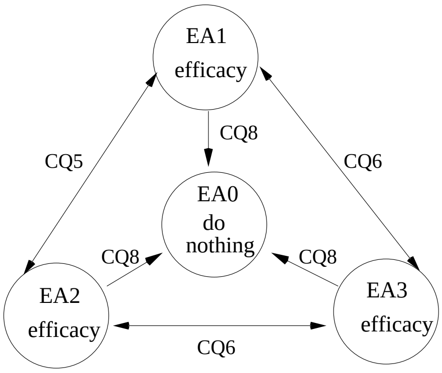


Figure 7.1: Initial argumentation framework.

The Policy Agent contains organisation specific information to determine preferences between goals, effects and actions. Any criteria could be used here. Although here no assumptions are made about the nature of the Policy Agent, it too could take the form of an argumentation agent like the Drama agent, and construct arguments for these preferences. For the purposes of the example it will be assumed that the Policy Agent prefers the effect ‘reduce platelet adhesion’ as a means by which the goal can be realised, since the effect of increasing blood clot dispersal agents has potentially more undesirable side effects. Hence, the Policy Agent will favour actions with the former effect over actions with the latter effect. This, however, does not discriminate between aspirin and chlopidogrel. Again many criteria are possible: it could depend on local stocks held, or a local preference for generic drugs. Here the assumption will be made that cost is the basis for preference and that aspirin is cheaper than chlopidogrel.

As used in the example from the previous chapter, these preferences are included in the argumentation framework by adding them as nodes blocking the attack of the arguments justifying the less preferred actions. The value given to the preferences is ‘choice’, which is always taken as the *least preferred value* in the framework. We thus get to Figure 7.2. Now EA1 (and the various goal and action choices) will form the preferred extension of this framework, and so this action is currently the best candidate. There remain, however, some further critical questions that can be asked of EA1.

EA1 assumed that aspirin was not contraindicated. CQ1 instructs us to test this assumption. This is the role of the *Safety Agent*. The Safety Agent has knowledge of contraindications of the various drugs, and the reasons for the contraindication. Again I take the Safety Agent to be in the form of a very simple Prolog based KBS which may contain:

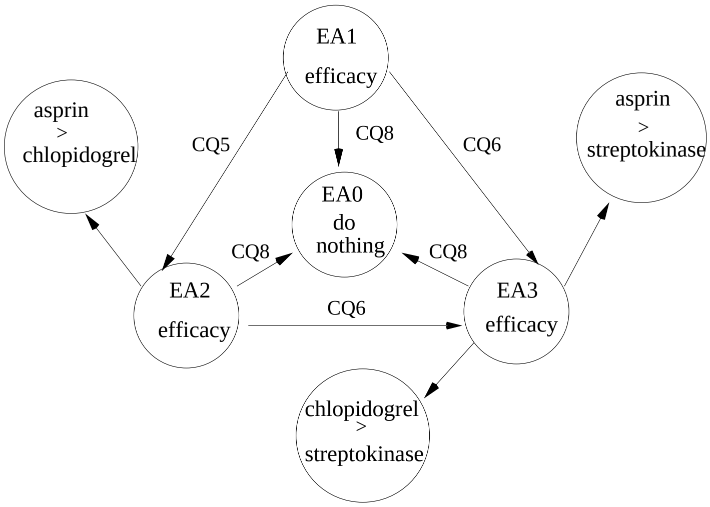


Figure 7.2: Argumentation framework with goal and action preferences.

```
contraindicated(X,Y):-
  risk_of_gastric_ulceration(X,Y).

risk_of_gastric_ulceration(X,Y):-
  increased_acidity(X,Y),
  history_of_gastritis(X),
  not acid_reducing_therapy(X).

increased_acidity(X,aspirin).
```

When contacted by the Drama agent it will use this knowledge, together with patient specific information obtained from the *Patient Agent* to inform the Drama agent that since the patient has a history of gastritis, aspirin is contraindicated because its acidity may result in gastric ulceration. The Drama agent will form this into an argument motivated by the value of ‘safety’. Note that because each of the information sources represents a particular perspective on the problem, the Drama agent may ascribe a motivating value to the argument on the basis of its source.

EA4 R4: Where there is a history of gastritis and no acid reducing therapy
A4: we should not prescribe aspirin
S4: so as not to cause excess acidity
G4: so as not to risk ulceration
V4: and so promotes the value of safety.

When EA4 is added to the argumentation framework, EA4 attacks EA1. Assuming that ‘safety’ is preferred to ‘efficacy’, EA4 defeats EA1 and so EA2 replaces EA1 in the preferred extension. The argumentation framework showing this is given in Figure 7.3.

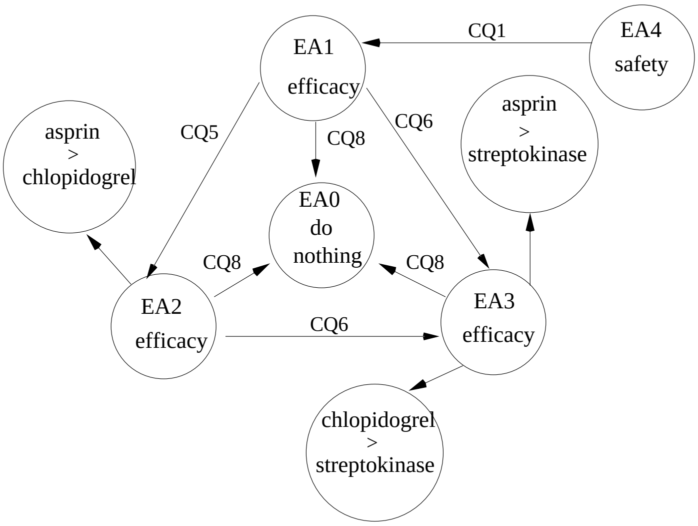


Figure 7.3: Argumentation framework with the addition of the attack of EA4.

Assuming EA2 cannot be attacked by CQ1, the next critique follows from CQ9. ‘Efficacy’ is not the only value: any action must be acceptable within the cost constraints of the organisation. Answering this critical question is the province of the *Cost Agent*. This agent will have knowledge of the budgetary constraints on treatment, and will compare the cost of the proposed treatment with these constraints. Suppose that chlopidogrel exceeds these limits. At the minimum this is simply a query as to whether the cost of the treatment exceeds a given threshold, posed to a database of treatment costs. The Drama agent can now form the argument EA5:

EA5 R5: Where cost of chlopidogrel is £N and budget $=$ £M and N $>$ M
A5: we should not prescribe chlopidogrel
S5: which would cost £N
G5: exceeding our budget
V5: which demotes the value of financial prudence.

The argumentation framework showing the addition of this argument is given in Figure 7.4.

Adding EA5 means that EA2 is defeated if ‘cost’ is preferred to ‘efficacy’. This still leaves EA3 unchallenged, and so we critique the proposal to prescribe streptoki-

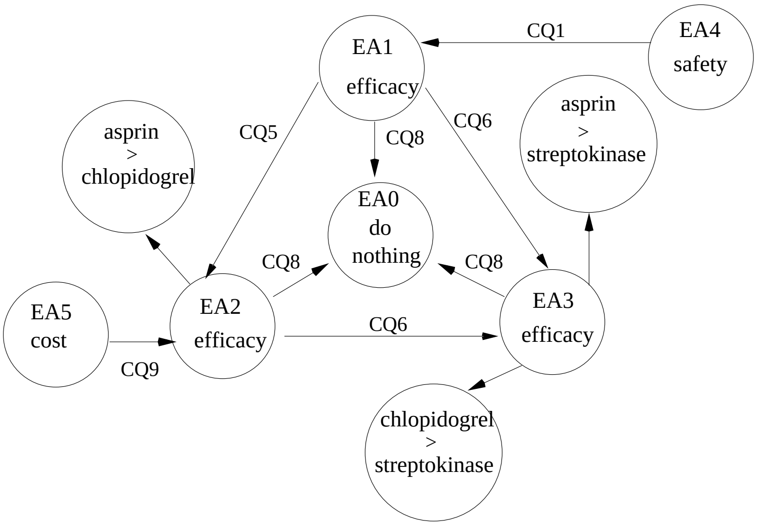


Figure 7.4: Argumentation framework with the addition of the attack of EA5.

nase, by returning to CQ1 and CQ9. Suppose that streptokinase is not contraindicated, and that it falls within the cost constraints. There remains CQ3, and we must now investigate whether streptokinase will be effective for the particular individual we are treating. The *Efficacy Agent* will contain specific data from clinical trials and past cases indicating the efficacy of actions with respect to treatment goals for particular patient groups. Perhaps (and this is simply an illustrative conjecture on my part) the efficacy of streptokinase has been found to depend on age. The Efficacy Agent may then contain information such as:

```
effectiveness(X, streptokinase, prevent_blood_clotting, 90):-
  age(X,A),A < 50.

effectiveness(X, streptokinase, prevent_blood_clotting, 30):-
  age(X,A),A > 49.

acceptable(X,Treatment, prevent_blood_clotting):-
  effectiveness(X,Treatment,E),E > 75.
```

Together with particular patient data obtained from the Patient Agent, the Efficacy Agent passes this information to the Drama agent which expresses it as EA6:

EA6 R6: Where patient is aged 72
A6: we should not prescribe streptokinase

S6: as the likelihood of success is 30%
G6: which is below the required threshold
V6: which demotes the value of efficacy.

The argumentation framework showing this is given in Figure 7.5.

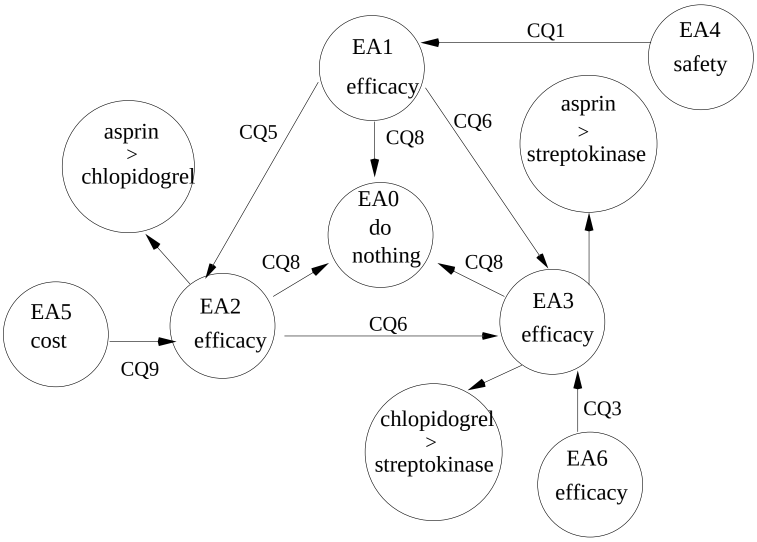


Figure 7.5: Argumentation framework with the addition of the attack of EA6.

Now EA3 is attacked by an argument with the same value and so is defeated. If ‘safety’ is preferred to ‘efficacy’ then EA1 is defeated by EA4. If ‘cost’ is preferred to ‘efficacy’ then EA2 is defeated by EA5. This would mean that EA0 would be included in the preferred extension as all its attackers are defeated. However, as stated from the outset, this is unacceptable as the patient’s health is then in jeopardy. There are two possibilities: either we must re-order our values so that ‘efficacy’ is preferred to one of ‘safety’ or ‘cost’, or else we must find an argument with which to defeat the attackers of one of EA1-3 and so reinstate one of our actions.

Suppose we re-order the values so as to prefer ‘efficacy’ to at least one of the other values i.e., we must choose whether we disregard ‘safety’ or ‘cost’. The choice will depend on the particular circumstances: it may be that the Drama agent is allowed to exceed budget if necessary, in which case ‘efficacy’ will be preferred to ‘cost’ and chlopidogrel will be prescribed. But if the cost constraint is rigid, there may be no better option than to disregard the contraindications and risk using aspirin, believing the complications to be less threatening than the immediate danger.

These hard choices can, however, be avoided if we can succeed in defeating one of the attacking arguments. We therefore run through our critical questions with respect to the arguments currently in the preferred extension of the framework. CQ1 can be posed with respect to EA4, as it is predicated on an assumption that in this situation there is no acid reducing therapy prescribed to the patient. We may therefore return to another Treatment Agent and attempt to find such an acid reducing therapy. This will supply the knowledge that a proton pump inhibitor (a particular type of acid reducing therapy) will have the desired effect. We can form this into EA7:

EA7 R7: Where there are no contraindications
A7: prescribing a proton pump inhibitor
S7: will prevent excess acidity
G7: removing risk of ulceration
V7: and so promotes the value of safety.

The argumentation framework showing the addition of the attack of EA7 is shown in Figure 7.6:

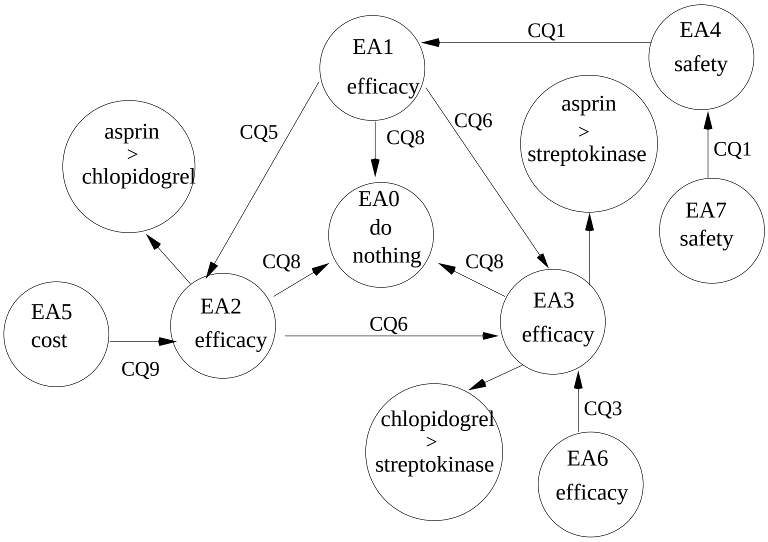


Figure 7.6: Final argumentation framework showing all critiques.

Of course, EA7 is now subject to the critical questions. Assuming, however, that there are no alternatives, that it is not contraindicated, within budget and likely to be effective, the argument gathering stops with Figure 7.6.

Figure 7.6 shows the full argumentation framework as we have now exhausted the relevant set of critical questions. The preferred extension is computed by first including the arguments with no attackers: EA5, EA6 and EA7. EA7 defeats EA4 because they both are motivated by the same value. This means that EA1 can be included, as its only attacker is defeated. EA1 thus defeats EA2 and EA3, again because they are motivated by the same value, and also excludes EA0, as desired. This in turn means that the three action preferences are no longer attacked and can be added to the preferred extension. Note that in this case we need express no value preferences: the preferred extension is the same irrespective of value order. From this we conclude that aspirin is the preferred treatment, as stated in EA1, and should be recognised as such by any audience.

7.4 Summary

In this chapter I have supplied a second domain to demonstrate how BDI agents can use my model of practical reasoning, with the specific example being to reason about the medical treatment of a patient. The example features a number of interesting elements that are sufficiently different from those revealed in the political example of the preceding chapter. A number of information sources provide the relevant data and the reasoning is all conducted by one central agent. The set of attacks that featured in this example focused on context dependant elements of the reasoning involved, providing proof that my method is effective in handling such an issue, as it was intended to do so. A more detailed discussion of the interesting features of this example will be given in Chapter 9.

Chapter 8

Application to Law

In this chapter I apply the definitions and methods described in Chapter 5 to a third example, taken from the legal domain, to show how BDI agents can use my model to reason about legal cases. I show how the reasoning in a well documented case from property law can be reconstructed in terms of my account using BDI agents. Section 8.1 discusses the background setting for the example and describes the particular legal case that is being modelled. Section 8.2 describes how the arguments used in the case can be represented and reasoned about in terms of my model. Section 8.3 provides a discussion of how the arguments generated in my example relate to the original opinions that were delivered in the real-life case. Section 8.4 concludes with a summary.

8.1 Background

One of the first projects in AI and Law, the TAXMAN project [112] of McCarty and Sridharan (most recently reported in [111]) had as its goal providing a computational means of generating the majority and minority opinions in a celebrated tax law case, *Eisner vs Macomber*, 252 U.S. 189 (1920). The work described in this chapter is in that tradition: here I will present a computational means of simulating the opinion and dissent in perhaps the most famous case in property law, *Pierson vs Post*, 3 Cai R 175 2 Am Dec 264 (Supreme Court of New York, 1805), said to have been read by (or at least assigned to) every law student in America. As a bonus I will also consider some additional arguments that have arisen in subsequent commentary and discussion.

The approach here will be to model the various participants in the debate as different agents. The disagreements in the case can be seen as grounded in divergent beliefs, goals and values, and therefore I will use different agents to represent the different views that can be brought to bear on the problem. In this example I will first recapitulate the details of the actual case, then I will show how the beliefs, desires and values of the four agents pertinent to the problem will be represented, according to the formalism described in Chapter 5. I then describe how the agents generate the arguments on the basis of their knowledge, and show the relations between these arguments as a set of VAFs. I end the example by relating this reconstruction to the opinions in the original decision.

I begin by giving a summary of the decision in *Pierson vs Post*.[1] The language used is appealingly extravagant and may in part account for the popularity of the case in teaching. It begins with a statement of the facts. After giving the procedural context the facts are stated as:

> "Post, being in possession of certain dogs and hounds under his command, did, upon a certain wild and uninhabited, unpossessed and waste land, called the beach, find and start one of those noxious beasts called a fox, and whilst there hunting, chasing and pursuing the same with his dogs and hounds, and when in view thereof, Pierson, well knowing the fox was so hunted and pursued, did, in the sight of Post, to prevent his catching the same, kill and carry it off. A verdict having been rendered for the plaintiff below, the defendant there sued out a certiorari and now assigned for error, that the declaration and the matters therein contained were not sufficient in law to maintain an action."

The opinion of the court was delivered by Tompkins, J. The decision can be seen as a sequences of parts, to which I will give identifying numbers Tn for later reference. He begins by stating the question to be determined (T1):

> "The question submitted by the counsel in this cause for our determination is, whether Lodowick Post, by the pursuit with his hounds in the manner alleged in his declaration, acquired such a right to, or property in, the fox, as will sustain an action against Pierson for killing and taking him away?"

The next paragraph (T2) discusses a number of authorities on the question of whether a wild animal can be owned other than through bodily possession, or at least mortal wounding. Tompkins concludes:

> "The foregoing authorities are decisive to show that mere pursuit gave Post no legal right to the fox, but that he became the property of Pierson, who intercepted and killed him."

He then (T3) dismisses a number of previous, mostly English, cases as irrelevant because they:

[1] The text of this decision is available on a number of websites e.g., http://www.saucyintruder.org/pages/pierson.html

> "... have either been discussed and decided upon the principles of their positive statute regulations, or have arisen between the huntsman and the owner of the land upon which beasts ferae naturae have been apprehended ..."

He next returns to his authorities (T4), and whilst being inclined to accept that wounding would constitute possession, states:

> "The case now under consideration is one of mere pursuit, and presents no circumstances or acts which can bright it within the definition of occupancy by Puffendorf, or Grotius, or the ideas of Barbeyrac upon that subject."

Next (T5) he considers a precedent case, *Keeble vs Hickeringill*, 11 East 574, 103 Eng Rep 1127 (Queen's Bench, 1707). This case had been cited as an example of where malicious interference in hunting was deemed to provide a reason for remedy. Tompkins distinguished this both on the grounds that Keeble suffered economic loss, and that the animals were on his own land:

> "... the action was for maliciously hindering and disturbing the plaintiff in the exercise and enjoyment of a private franchise; in the report of the same case, (3 Salk. 9) Holt, Ch. J., states, that the ducks were in the plaintiff's decoy pond, and so in his possession, from which it is obvious the court laid much stress in their opinion upon the plaintiff's possession of the ducks, ratione soli."

He then (T6) motivates his decision by a desire that the law should be clear:

> "We are the more readily inclined to confine possession or occupancy of beasts ferae naturae, within the limits prescribed by the learned authors above cited, for the sake of certainty, and preserving peace and order in society. If the first seeing, starting, or pursuing such animals, without having so wounded, circumvented or ensnared them, so as to deprive them of their natural liberty, and subject them to the control of their pursuer, should afford the basis of actions against others for intercepting and killing them, it would prove a fertile source of quarrels and litigation."

Finally (T7) he concludes by saying that even if any malice was involved this "act was productive of no injury or damage from which a legal remedy can be applied.", suggesting that such damage needs to be economic to provide any remedy: the law cannot compensate for loss of sport.

The overall thrust of this decision seems to be that the law is rather clear as it stands: the only question is ownership, and that ownership in a wild animal cannot be acquired through mere pursuit. Moreover, where there is no measurable damage, no legal remedy is appropriate.

Livingston, J. then gives his dissent. Again, I number its parts Ln for later reference. He (L1) agrees that there is a single question: whether pursuit of the fox gave "such an interest in the animal, as to have a right of action against another". He then says (L2) that such cases should not be brought to court but arbitrated by sportsmen (ignoring any concerns of natural justice that would result from the bias in favour of Post at such a tribunal). He then (L3) argues that hunting should be encouraged as the depredation of foxes "on farmers and on barn yards have not been forgotten; and to put him to death wherever found, is allowed to be meritorious, and of public benefit" and that no one would hunt if their sport were regularly spoiled by interventions such as that of Pierson. He says (L4) the authorities cited are old, and that the court is able to state a new law: "if men themselves change with the times, why should not laws also undergo an alteration?" In any event the authorities do not require bodily possession and so a finding for Post would be compatible with them. The crux of his argument (L5) is that

> "... the interest of our husbandmen, the most useful of men in any community, will be advanced by the destruction of a beast so pernicious and incorrigible, we cannot greatly err, in saying, that a pursuit like the present, through waste and unoccupied lands, and which must inevitably and speedily have terminated in corporal possession, or bodily seisin, confers such a right to the object of it, as to make any one a wrongdoer, who shall interfere and shoulder the spoil."

In sum: since fox hunting is of public benefit because it assists farmers it should be encouraged by giving the sportsman protection of the law.

The arguments of Tompkins and Livingston are couched in very different terms. Whereas Tompkins confines himself to discussion in terms of legal concepts – which to him clearly provide no basis for remedy, Livingston talks mainly about the real world, and whether fox hunting is desirable or not, and argues that if it is, the legal concepts should be interpreted so as to provide a remedy.

In the reconstruction of the arguments two different agents to represent Tompkins and Livingston will be used. These agents will be referred to as T and L respectively. Two additional agents will also be used to make points not raised in the decision, but which have emerged in subsequent debate.

The first of these additional agents disputes Livingston's claims about the benefit of hunting. In the novels of Anthony Trollope the topic of fox hunting features quite prominently. In one of his novels, *The American Senator* [158], a major sub-plot concerns a farmer who poisons a fox. This outrages the hunting community, since they wish to preserve foxes for their sport. It is quite clear that Trollope, who is a fervent pro-hunter, recognises that but for hunting, foxes would be rapidly eliminated by farmers through the more efficient pest control methods of snaring, poisoning, gassing and shooting. If we agree with Livingston that it is "meritorious" and "of public benefit" to "put [a fox] to death wherever found", then hunting should be discouraged, since where hunting is encouraged these more efficient methods are subject to social stigma. Trollope, however, does wish to encourage hunting on its intrinsic merits: he would therefore wish the law to condemn Pierson's malicious interference in the sport. I shall call this agent A, for Anthony.

The final agent also disputes whether hunting should be encouraged. A recent Act of Parliament means that fox hunting in the traditional manner is now illegal in the UK. The argument here has been solely based on the cruelty of hunting: shooting and gassing are preferred on grounds of humaneness rather than efficiency. On such a view Pierson is acting in a laudable manner, by saving the fox pain, and it is the actions that discourage hunting that should be encouraged. I will call an agent with such a view agent B, after Tony Banks, MP, who was a vocal opponent of hunting during this debate.

In the next section I will instantiate these four agents with the appropriate beliefs, desires and values, in accordance with the definitions given in Chapter 5.

8.2 Generating The Arguments

I begin by identifying desires and values. From Definition 4 we need to identify a set of desires for the agents, and give conditions under which the agents will accept that these desires are realised. Definition 5 requires us to associate these desires with a value, and an indication of whether the desire promotes or demotes the value. Tables 8.1 and 8.2 list the set of desires, conditions, values and degrees that will be used. Since varying degrees of promotion are not considered in this example, it represents the promotion of values as 1 where the value is promoted and -1 where it is demoted. Table 8.1 gives the initial desires, and Table 8.2 those that may be derived in the course of the debate. So for example, the desire for "clear law" is satisfied either if the case is decided for the plaintiff when ownership is established, or it is decided for the defendant when no ownership can be established. Similarly, restricted trade requires a decision for the defendant, malicious intent to be found and for the plaintiff to be engaged in productive activity.

Table 8.1: **Possible desires and values in the initial situation**

No.	Desire	Value affected	Condition to satisfy
1	Clear Law, promotes	Less Litigation	Ownership, Plaintiff OR No ownership, Defendant OR No ownership, No possession
2	Unclear Law, demotes	Less Litigation	Ownership, No Possession. No Ownership, Plaintiff. No Ownership, Possession.
3	Trade Restricted, demotes	Economic Benefit	Malicious Intent, Productive Activity, Defendant
4	Malice Condemned, promotes	Public Benefit	Malicious Intent, Plaintiff
5	Malice Condoned, demotes	Public Benefit	Malicious Intent, Defendant
6	Less Threat to Others, promotes	Public Benefit	Fewer Foxes, Farmers Protected
7	More Threat to Others, demotes	Public Benefit	¬Fewer Foxes, ¬Farmers Protected
8	More Suffering, demotes	Humaneness	¬Reduced Animal Suffering
9	Less Suffering, promotes	Humaneness	Reduced Animal Suffering

Table 8.2: **Derivable desires and values**

No.	Desire	Value Affected	Condition to satisfy
10	Hunting Encouraged, promotes	Public Benefit	Ownership, Pursuit
11	Hunting Discouraged, promotes	Humaneness	¬Pursuit, No ownership

There are 4 agents in the situation: Livingston(L), Tompkins(T), Banks(B) and Trollope(A). Each agent has different desires they wish to achieve and has different values they wish to promote, though many of these will be in common. From Table 8.1 all agents ascribe to desires 1 and 2 and 3. Agents T, L and A do not accept desires 8 and 9 as they do not regard 'reducing animal suffering' as promoting 'humaneness', animal suffering not being a consideration in their pre-animal rights way of thinking. Additionally, agent T does not accept desires 4 to 7 as he does not regard 'public benefit' as a value which the law should recognise. The agents may also adopt the derived desires in the course of their reasoning.

Seven propositions about the world are used to describe the given situation and these are as follows:

- F1: Post was in pursuit of the fox.
- F2: Post had neither captured nor wounded the fox (he had no possession of the fox).
- F3: Pierson killed the fox to spoil Post's sport (Pierson had malicious intent).
- F4: Foxes kill livestock.
- F5: Encouraging hunting will reduce the number of foxes.
- F6: Reducing the number of foxes protects the livestock of farmers.
- F7: If hunting is discouraged, needless animal suffering is not inflicted.

The agents differ quite widely as to the facts. Each agent ascribes to these propositions as shown in Table 8.3 with 1 representing belief in the proposition, -1 representing disbelief in the proposition and 0 representing unknown to show that the agent has subscribed to neither belief nor disbelief in the proposition.

Table 8.3: **Propositions about the world**

Agent	F1	F2	F3	F4	F5	F6	F7
L	1	1	0	1	1	1	0
T	1	1	1	1	0	1	0
B	1	1	-1	0	0	0	1
A	1	1	1	1	-1	-1	0

Based upon the beliefs and desires given in the above tables, each agent can provide one or more instantiation of AS1. The figures presented below give three argumentation frameworks to show the views of the agents at three different levels: the level of facts about the world, at which new desires can be derived; the level at which the legal system connects with the world to achieve these desires, and at the level of pure legal concepts. These levels are familiar from other work in AI and Law, and are explicit in the functional ontology of Valente [160], and some discussions of expert systems within the logic programming paradigm, such as [22]. Conclusions at lower levels will be used as premises at higher levels. The emergence of these levels of reasoning will be discussed further in Chapter 9 in the evaluation of the example applications.

Each of the argumentation frameworks will first be presented, followed by the instantiations of AS1 and then any attacks that can be made on these instantiations by satisfying the appropriate pre-conditions, as described in Chapter 5. In the figures, nodes represent arguments. They are labelled with an identifier, the associated value, if any[2], and on the right hand side, the agent(s) introducing the argument. Arcs are labelled with the number of the attack they represent. I then summarise what can be deduced from the framework in order to proceed to the next level in the argument. Below, in Figure 8.1, is the argumentation framework for Level 1 schemes:

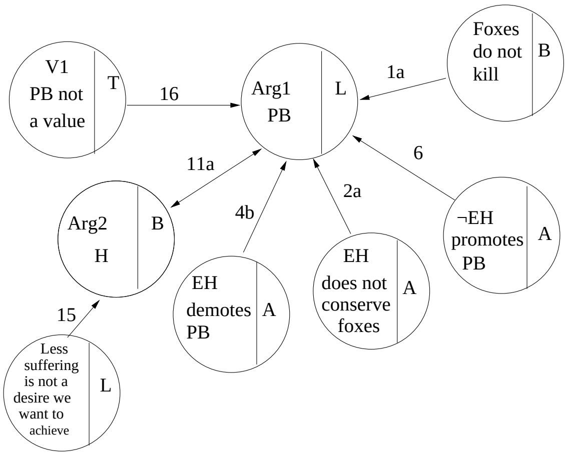


Figure 8.1: Level 1: Arguments about the world.

This argumentation framework is constructed from the following arguments. In expressing the arguments in this example I have, for readability reasons, avoided repetition of facts by using only the relevant subset of R. This means that S can be omitted altogether as G comprises this relevant subset, together with the decision taken. S is of importance only if the results of the action are in doubt and here the action cannot fail

[2] Where no value is given, the argument is a statement of fact, and so can be taken as having ‘truth’ as its value. In my approach, and following [26], ‘truth’ is the most highly ranked value for all audiences.

as it is simply decided that something will be regarded as true, for example, that the plaintiff wins. It could be that there are some other arguments relevant to this reconstruction that are not included here that require the use of S, then it would need to be included.

Arg1

R1: Where foxes kill livestock, encouraging hunting leads to fewer foxes and fewer foxes means farmers are protected
A1: encourage fox hunting
G1: as fewer foxes and farmers protected
V1: promotes public benefit.

Agent L puts forward Arg1 and this is attacked by Arg2 using attack 11a which is put forward by agent B:

Arg2

R2: Where fox hunting is cruel
A2: discourage fox hunting
G2: as reduced animal suffering
V2: promotes humaneness.

This argument is mutually attacked by agent L's original statement made in Arg1 but agent L can also attack it using attack 15 which states that L does not believe that 'reduced animal suffering' is a desire that we want to achieve. Agent B can also make a second attack by disputing the fact 'foxes kill' using attack 1a. Agent A can also attack agent L's Arg1 by using 3 different attacks; attack 4b, 2a or 6. Using attack 4b, agent A states his belief that encouraging hunting will demote public benefit, advanced by Livingston, as hunters preserve foxes and destroy crops, both of which are not of benefit to the public. Using attack 2a he further states that contrary to Livingston's belief that there will be fewer foxes if hunting is encouraged, there will in fact be more, as hunters preserve foxes for their sport. Using attack 6 agent A then states that there is an alternative action of not encouraging hunting, as then foxes will be wiped out by other means of pest control, and fewer foxes promotes 'public benefit'. Finally, agent T can make attack 16 on Arg1 by stating that 'public benefit' is not a value we should be trying to promote.

From the argumentation framework in Figure 8.1 agent L can, by choosing to rank 'public benefit' over 'humaneness', deduce that hunting should be encouraged and agents B and A can deduce that hunting should be discouraged, using their own preferences. T, by accepting L's argument against Arg2, need subscribe to neither argument, and so derives no additional desires from this level of the debate.

We can now move on to the next level, giving the argumentation framework shown below in Figure 8.2:

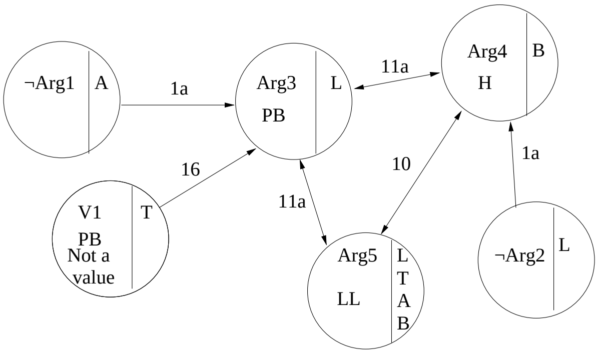


Figure 8.2: Level 2: Linking to legal concepts.

This argumentation framework is constructed from the following instantiations of AS1:

Arg3

R3: Where there is pursuit and fox hunting is to be encouraged
A3: find ownership
G3: as hunting encouraged
V3: promotes public benefit.

Arg4

R4: Where there is pursuit and fox hunting is to be discouraged
A4: find no ownership
G4: as hunting discouraged
V4: promotes humaneness.

Arg5

R5: Where there is no possession
A5: find no ownership
G5: as finding no ownership where no possession
V5: promotes less litigation.

Agent L puts forward Arg3. Firstly, this is attacked by agent A using attack 1a, stating that he does not believe Arg1 from the previous framework to be true. Agent T also attacks Arg3 by using attack 16 which states that ‘public benefit’ is not a value. There is then a 3-cycle of attacks: all agents using attack 11a to attack Arg3 with Arg5. This is itself attacked by Arg3. The next attack in the cycle is also a mutual one put forward by agent B using attack 10 to state Arg4. Arg4 also mutually attacks Arg3 using attack 11a, which completes the 3-cycle. However, Arg4 is attacked by agent L using attack 1a stating that he does not believe that Arg2 from the previous framework holds.

Figure 8.2 debates whether or not ownership is to be attributed on these facts. L uses Arg3 to say that ownership should be attributed, relying on his view of what the facts about foxes are from Level 1. L uses a preference for ‘public benefit’ over ‘less litigation’ to avoid defeat by Arg5. He attacks Arg4, the favoured argument of B, because he does not accept Arg2 from Level 1, since ‘humaneness’ is not among his values. The attack of A, using attack 1a, can be ignored by L as it turned on a factual disagreement in the previous level. All except L agree that Arg3 is defeated, although for different reasons, and so accept Arg5. L accepts Arg5, but believes that its force is insufficient to defeat Arg3. We can now move on to the top level arguments, giving the argumentation framework shown below in Figure 8.3:

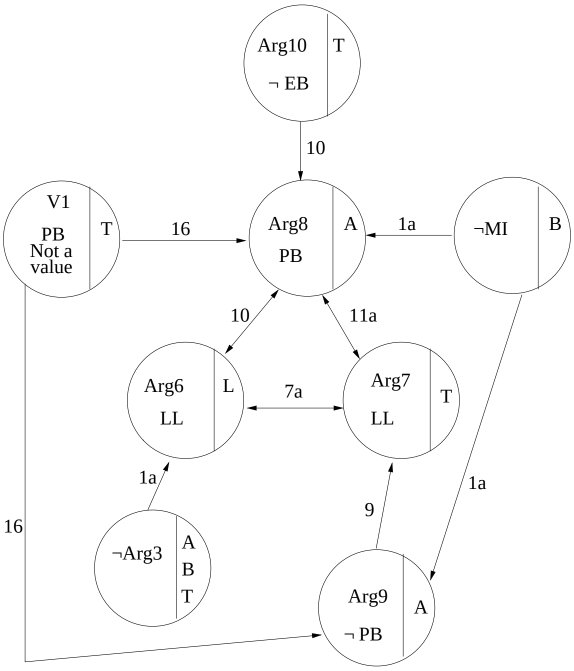


Figure 8.3: Level 3: Arguments in terms of legal concepts.

This argumentation framework is constructed from the following argument schemes:

Arg6

R6: Where there is ownership
A6: find for plaintiff
G6: as finding for plaintiff with ownership
V6: promotes less litigation.

Arg7

R7: Where there is no ownership
A7: find for defendant
G7: as finding for defendant where there is no ownership
V7: promotes less litigation.

Arg8

R8: Where there is malicious interference by defendant
A8: find for plaintiff
G8: as finding for plaintiff where there is malicious interference condemns immoral behaviour
V8: promotes public benefit.

Arg9

R9: Where there is malicious interference by defendant
A9: do not find for defendant
G9: as finding for defendant where there is malicious interference condones immoral behaviour
V9: demotes public benefit.

Arg10

R10: Given the facts of Keeble
A10: do not find for defendant
G10: as not finding for defendant where there is malicious interference and productive activity
V10: promotes economic benefit.

Agent L puts forward Arg6. This is immediately attacked by all of T, B and A who, for their different reasons, did not accept Arg3 from the previous framework and therefore deny its premise. Next, agent A, who wishes to find for Post not on grounds of the protection of farmers but to condemn Pierson's interference in Post's sport, uses attack 10 to state Arg8 (which is mutually attacked by Arg6), but this is in turn attacked by agent T's attack 16 and also by agent B's attack 1a stating that he does not believe that the interference was malicious. Agent T can make attacks 11a and 7a on Arg8 and Arg6 respectively by stating Arg7, which is his main argument, based on his acceptance of Arg5 at the previous level. This creates a 3-cycle of attacks between schemes Arg6, Arg7 and Arg8. However, Arg7 is then attacked by agent A using attack 9 which states Arg9. Like Arg8 this can be attacked by the two existing nodes in which agent T uses attack 16 to say 'public benefit' is not a value and agent B uses attack 1a to state that he does not believe that there was malicious interference.

The one precedent case explicitly cited in the decision is *Keeble*. The role of *Keeble* is to provide support for Arg8, in the manner described in [75]. This, however, can be attacked using attack 10 as described in that paper, as the finding for the plaintiff can be motivated either by desire 3 from Table 8.1, as Keeble was engaged in a profitable enterprise, or by a desire expressing protection of property rights. These alternative interpretations of *Keeble*, could be added to the framework: in Figure 8.3 the first of these is added as Arg10, making attack 10. The second challenge is not represented here as no beliefs or desires relating to property have been considered here.

This now completes the final framework and so we can deduce whether the plaintiff has remedy or not. L, who accepts Arg3, and gives prime importance to 'public benefit' will use Arg6 to determine his decision. A, who also gives primacy to 'public benefit', but rejects the facts on which Arg6 is ultimately based will use Arg8. B rejects the premises of both Arg6 and Arg8 and so he finds Arg7 acceptable. Finally, T accepts Arg7 as it is the only argument grounded on a value of which (in his opinion) the law should take note.

8.3 Relating the Arguments and the Opinions

In this section I return to the opinions of the actual case summarised in Section 8.1, and relate them to the various components of the argumentation frameworks produced in the previous section. I begin by relating the arguments and attacks put forward by T and L in these frameworks to the opinions of Tompkins and Livingston. The arguments of A at Level 3 will also need to be considered, since these are referred to by Tompkins in order to be rejected. At Level 3, A can be seen as the representative of the hunting aficionado, and his arguments reflect those that we might expect Post, or his counsel, to advance. I do not expect to be able to reflect the structure of the opinions, nor, of course, the extraordinary language used to deliver them, but I do hope to identify the reasoning elements corresponding to T1-8 and L1-5.

I will begin by considering the opinion of Tompkins. I will proceed top down, as this corresponds most closely to the structure of that opinion. Therefore, consider first Figure 8.3. Tompkins must primarily dispose of alleged precedent cases, represented in Figure 8.3 by Arg8. He first dismisses a number of cases as irrelevant (T3) and distinguishes *Keeble* (T5). Since we do not know the cases referred to in T3 or which argument they were supposed to support they have not been represented here. The attacks on the interpretation of *Keeble* in T5 are represented by Arg10. Once the alleged precedent has been dismissed, Arg8 can be eliminated by denying that its value is a proper concern (V1 in Figure 3). This corresponds to T7, “no injury or damage for which a legal remedy can be applied”. With Arg8 eliminated, the question turns on whether the premises of Arg6 or Arg7 are accepted. This is the question expressed in T1, and which is answered in the framework of Figure 8.2.

In the framework at Level 2, T adopts Arg5, that the law is clear that where there is no possession there should be no ownership. That the law is clear on this point and that there are neither cases nor authorities to suggest that pursuit of a wild animal may constitute ownership, is the point of T2 in Tompkins’ opinion. The conclusion that mere pursuit cannot count as ownership is explicitly expressed at the end of T4. The purpose motivating Arg5 is expressed in T6, “for the sake of certainty”.

This relates all seven components of Tompkins’ opinion to the agent-based account given here. T1 states the choice to be made at Level 3, T2 and T4 provide the factual basis of Arg5, which is then motivated by the value supplied in T6, and which leads to a denial of a premise in Arg6. T3, T5 and T7 remove unfavoured arguments from Level 3. Tompkins, like agent T, has no need to descend to the issues raised at Level 1.

Turning now to Livingston, he first agrees with Tompkins’ view of Level 3 (L1). L2 seems to endorse the opinion that the case should never have come before a court, and suggests that in a tribunal of sportsmen, Arg8 would be followed. In a court, however, clarity of law is important, so this can be taken as acceptance of the force of Arg5, which is motivated by a desire to reduce the potential for these matters to be litigated. Livingston’s main argument is Arg1, stated and motivated in L3, to kill foxes “is allowed to be meritorious, and of public benefit”, the value being expressed again at L5. L4 is concerned to argue that the court may make the law (because only if this is so are Level 1 concerns able to be introduced). It is this which expresses the preference of ‘public benefit’ over ‘less litigation’, which is necessary if Arg3 is to succeed over Arg5. He suggests that the public benefit was not recognised by Justinian law (the original authority cited to establish rights of possession over wild animals) only because fox hunting was not then in fashion. Had it been, “the lawyers who composed his institutes would have taken care not to pass it by, without suitable encouragement”.

Tables 8.4 and 8.5 summarise this discussion by listing the components of the argumentation framework, together with the agents that introduced them, and the section of the opinion which they represent. These tables show that each of the sections of the opinion can be linked to a component in the framework, with the exception of T3 and L2. T3 is similar to T5, but is omitted because, unlike *Keeble*, we do not have sufficient information about the cases dismissed as irrelevant to represent them. L2 is something of an aside, expressing sympathy with Arg8, whilst recognising that it cannot prevail over Arg7 in a court of law since for L, consideration of public benefit is proper to Level 2.

Table 8.4: **Arguments introduced (or mentioned if starred) in opinions**

Argument	Agent	Opinion Section
Arg10	T	T5
Arg9	A	T7*
Arg8	A	T5*
Arg7	T	T1
Arg6	L	L1
Arg5	T, L, A, B	T2, T6
Arg3	L	L4
¬Arg3	T, A, B	T4
Arg1	L	L3, L5
Value 1 (V1)	T	T7

Table 8.5: **Attacks made (or mentioned if starred) in opinions**

Attack	Attacker	Attacked	Agent	Opinion Section
11a	Arg8	Arg7	A	T7*
7a	Arg7	Arg6	T	T1
1a	¬Arg3	Arg6	T, A, B	T4
16	V1	Arg8	T	T7
10	Arg10	Arg8	T	T5
11a	Arg3	Arg5	L	L4

Of the arguments in the frameworks, ¬MI (‘no malicious interference’) from Level 3 is omitted, as are Arg4 and ¬Arg2 from Level 2 and all except Arg1 from Level 1. ¬MI and Arg4 are proposed only by B and so represent a point of view which emerged after the decision. Similarly ¬Arg2, although attributed to L, appears only in order to attack Arg4. At Level 1, in the actual case no challenge was made to Livingston: again the arguments reflect later discussions. Thus all the reasoning moves in the framework that address concerns that arose at the time of the case are reflected in the opinions.

Note that Tompkins confines his considerations to Levels 2 and 3, whereas Livingston, who needs to argue instrumentally, must start at Level 1 with a discussion of the way of the world. The other two agents operate mainly at Levels 1 and 3, reflecting the fact that they are not producing essentially legal arguments. Agent A disagrees with agent L about the facts of the world (not the facts of the case), suggesting that fox hunting does nothing to reduce the fox population. At Level 1 agent A argues for a sense of what is fair over the legal question which Tompkins addresses. Agent B argues from a moral rather than a legal perspective, using a general moral value at Level 1 and a coloured interpretation of the facts of the case at Level 3.

This concludes the analysis of the representation of this particular legal case in terms of my BDI agent account. Some further discussion of the levels of legal reasoning involved in this reconstruction can be found in [8].

8.4 Summary

The example presented in this chapter has shown that the BDI application of my theory of persuasion over action can be used to reconstruct the reasoning in a well known legal case. The reconstruction has provided a useful domain to exercise my proposal of representing such debates as a multi-agent system with the different agents representing divergent beliefs, desires and values. In addition to the successful reconstruction of the case, this example has also revealed some interesting insights about the layered nature of the reasoning involved here. I will say more about this feature in my discussion of all the examples, which is given in the next chapter.

Chapter 9

Discussion of the Applications

In this chapter I evaluate the three examples from the previous chapters which make use of my approach to practical reasoning with value enhanced BDI agents. This begins in Section 9.1 which discusses the political example using BDI agents and the interesting features that emerged from the example with regard to the notion of accrual of arguments. This evaluation also provides a discussion of how the example relates to the PARMENIDES system discussed earlier on in Chapter 6. Section 9.2 discusses the medical example and the merits of treating the reasoning involved in terms of a distributed system. Section 9.3 provides an evaluation of the legal example and discusses an interesting feature that emerged from the example regarding the different levels of reasoning involved. Section 9.4 provides an evaluation of the general underlying approach and the conclusions that can be drawn from its application to these three theoretical examples. Section 9.5 concludes with a summary.

9.1 Evaluation of the Political Application

The first example application I presented to illustrate my approach to representation of practical reasoning in BDI agents was the political application of Chapter 6. The example used different BDI agents to model a recent political debate involving the Government's justification of a proposed action and its subsequent scrutiny by members of Parliament and the public. Below I discuss the points of interest arising from this example.

One of the most distinctive features to emerge from this application is that it requires the modelling of the extent to which the values involved are promoted. Although the example did not represent all the attacks that the arguments were subjected to in the real-life scenario, I chose a relevant subset of the most prominent critiques that were debated in Parliament at the time. Looking at these critiques, many of the arguments put forward were based upon the individual agents' beliefs as to whether or not the justifications for actions presented promoted or demoted their preferred value in some way or another. Thus, the representation of values has proved to be a crucial element in modelling this debate. As discussed in Chapter 2, it is not always the case that agreement will result between rational agents, even when they agree on all the facts of a situation. Recall, as Searle noted, “rational agents are likely to have different and inconsistent values and interests, each of which may be rationally acceptable.” [146, p. xv]. This observation has proved to be particularly pertinent to the example debate represented in this particular domain and application. As the example in Chapter 6 concluded, the action to be taken was dependant upon the individual agents’ preference orderings on values. In this way it is possible to see how disagreement occurred and why. This reflects the situation in real-life, as fiercely contested issues sometimes cannot be resolved and discussions have to terminate with the participants ‘agreeing to disagree’. This is true of the real-life debate on the invasion of Iraq. Thus, I believe that my model of practical reasoning is able to realistically represent such real-life debates on matters of practical action, whilst being able to explain how and why rational disagreement can and does occur in such situations.

Another feature of this application which I broached in Chapter 6 was the notion of accrual and strength of argument. The model of practical reasoning I have proposed allows for differing degrees of belief in the elements of a justification for action to be represented. This enables the strength of individual arguments to be assessed. As commented on in the example, people often try to lend further support to an argument by supplementing it with further justifications for why it should be carried out. Thus arguments can lend strength to one another, which presents a more compelling case for performing an action and can increase the persuasive power of the proponent of such arguments. These insights came to light when considering how this debate unfolded in real-life and how this could be represented by my account. Thus, it was not my original intention to try and model this concept but I have provided in Chapter 6 a brief discussion and preliminary outline for how this could be dealt with in my account. The mechanisms I discussed there to implement degrees of promotion and strengths of justification have been simple and preliminary, and are intended as an addition to the main purpose of the model. Nonetheless, they do illustrate in a minimal fashion how this model can capture the useful notion of accrual, and can distinguish different types of accrual. However, I do believe this outline could be extended to give further insight into how to deal with this notion of arguments being combined to build up strength for a particular point. There is other existing literature on this topic, e.g., from the legal argumentation field [129] and the uncertainty in AI community [119, 149, 171], which could be explored and this would provide an interesting and useful extension for future work.

I now discuss and compare the BDI agent political example of Chapter 6 with the PARMENIDES system detailed in the same chapter. The domain and debate used in these two examples is the same i.e., justification for the invasion of Iraq.

Firstly, PARMENIDES makes use of a subset of the attacks from the theory of persuasion over action, as does the example in Chapter 6. These two subsets of attacks are extremely similar with the exception of three additional attacks being used in PARMENIDES (attacks 7, 11 and 16). The purpose of PARMENIDES was to build a system in which users could critique a justification of an action in a particular domain and express their own views in the most complete way possible, using the theory. However, it would be perfectly acceptable to use the extra attacks found in PARMENIDES in Chapter 6's BDI agent example: they do not in fact arise because the example is limited to make use of only seven agents, whose beliefs and desires do not happen to satisfy the pre-conditions for these attacks. Thus, it would be possible to reconstruct all the arguments made in PARMENIDES in the format presented in Chapter 6 for reasoning with BDI agents, though this was not the original aim of the exercise.

One of the main motivations of PARMENIDES was to provide a system which facilitated debate between the Government and members of the public whilst being grounded in a firm model of argument that was transparent to the user. Conversely, the BDI model presented in Chapter 6 is intended for use solely by autonomous computer agents. However, I believe that there may be a useful link between the two models. As mentioned in the discussion of PARMENIDES in Chapter 6, all the information entered into the PARMENIDES system is stored in a back-end database. Therefore, it would be possible to reconstruct new positions on the issue from the users' responses by introducing agents to represent their views. These new positions could then be used as input to generate presumptive arguments to be used by BDI agents, as demonstrated in the BDI agent example of Chapter 6. This would allow for the gathering of a wide range of differing views on the topic and evaluation of the warrant of each view. As part of the practical reasoning process this would ensure that all possible scenarios have been considered and thus aid us in choosing the best action and justification for the issue in question. This is would be an interesting and seemingly feasible path to pursue between the two systems and is something that could be the focus of future work.

9.2 Evaluation of the Medical Application

The second example application I presented to illustrate my approach to the representation of practical reasoning in BDI agents was the medical application of Chapter 7. The representation of the scenario was presented as a distributed system for deliberative reasoning about the treatment of a patient and it has a has a number of worthwhile features, which I summarise below.

The account of practical reasoning that I have given in this thesis has been developed to capture features of this reasoning observed in the philosophical and informal logic literature, as discussed in Chapter 2. Some of these features have proved to be of particular importance in the application to the medical domain. The agents bring different information, representing different areas of medical knowledge and facts about individuals, and also represent different policies and perspectives (such as cost, efficacy, safety, etc.) relevant to the problem. This highlights the importance of viewing information from different perspectives (or audiences) with the different agents each bringing their own values to the problem. All these perspectives that need to be considered when making a medical decision are kept separate, and thus it is made explicit from which perspective the various arguments derive. This means that the perspectives can be given their due weight, but discounted if necessary. It is the job of the Drama agent to then perform the argumentation by coordinating these contributions and forming them into arguments to come to a decision.

The critical point of the reasoning in this example concerned the uncertainly of effects of different treatments. As individuals respond differently to the same treatment, the coordinating agent must gather as much information as possible about the patient and the likely effects of drugs on them. Such cases are obviously highly context dependant as individuals' health status and responsiveness to treatment is dependant upon a wide variety of factors. Thus any ordering of preferences must take the specific context into account. I believe that my account of practical reasoning contributes to meeting this objective.

Another worthwhile feature of this particular application is that the model is effected inside a single agent: the other agents in the system can therefore be conventional knowledge and database systems, simplifying their participation in other systems. If, however, more sophisticated resources are available, these can be used by the Drama agent without modification. Each of the information sources used by the central agent are dedicated to the provision of particular information, need not consider every eventuality, and play no part in the evaluation. This simplifies their construction and facilitates their reuse in other applications. Additionally, distinction can be made between information sources which are generic and those which are particular to a specific organisation or individual.

The final point to note is that the critiques that are posed against the putative solutions are made only as and when they can affect the dialectical status of arguments already advanced. This means that all reasoning undertaken is of potential relevance to the solution.

In summary, the approach to reasoning about decisions based on several information sources (such as is the case in medicine) using my model of practical reasoning has shown considerable potential in facilitating agents' decision making capabilities in such a context. Two particular benefits of the approach have been the use wherever possible of conventional and generic components, and the ability to make flexible and context dependent decisions, taking a number of perspectives into account.

9.3 Evaluation of the Legal Application

The third and final example application I presented to illustrate my approach to representation of practical reasoning in BDI agents was the legal application of Chapter 8.

The particular legal case chosen for the example is one that has been studied extensively and is well known in legal reasoning, as well as in AI and law [29, 32, 128, 143]. My intention was to demonstrate that real-life examples can be constructed in terms of my model of practical reasoning through the use of BDI agents to represent the actual beliefs, desires and values that were present in the case. I believe that this example has been successful in achieving this objective. In addition to this accomplishment, there are some other insights that can be drawn from my example, which I discuss below.

One interesting feature to emerge from the reconstruction of the debate is the layered nature of the process: instrumental arguments about the world provide premises for arguments about the application of legal concepts, which in turn form the basis for the resolution of the legal questions relating to the case.

In the example, the uppermost layer (Level 3) was concerned with legal concepts and the rights they conferred. In the particular opinions, disagreement at this level was based solely on whether or not ownership of the fox was ascribed: once this point was decided the consequences were clear and there were no conflicting considerations. The second layer (Level 2) concerned the ascription of these legal concepts, given the particular facts of the case under consideration. Here arguments for and against ascription of the legal concepts can come either from precedents (although there were no applicable precedents in the particular case), or from purposes derived from reasoning in the bottom layer (Level 1). At Level 1 people reason about the world in order to determine what the law *should* be, and conclusions from this level are used at Level 2. These three levels of reasoning are shown diagrammatically in Figure 9.1.

These levels constantly appear in work on case based reasoning in AI and Law. For example, the need to make the transition from facts about the world to legal concepts was a major concern of expert systems designers in the 1990s. The work of Breuker and his group (e.g., [39]), explored the need to represent knowledge of the world, knowledge of legal concepts and the connections between them. Work on legal expert systems in the logic programming tradition (e.g., [23]), tended to begin with a definition of the terms of the legislation, and then unpack the definitions of these terms using sufficient conditions expressed in factual terms taken from case law and expert guidance. These strands of work arose from a response to practical problems of design-

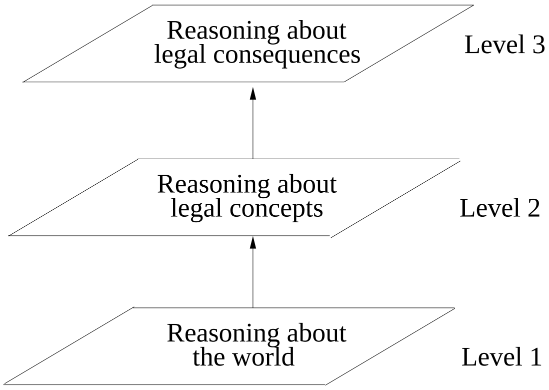

Reasoning about legal consequences
Level 3
Reasoning about legal concepts
Level 2
Reasoning about the world
Level 1

Figure 9.1: The three levels of legal reasoning emerging from the reconstruction of *Pierson vs Post*.

ing and building legal expert systems. A recent paper [99], gives a formal expression to these notions and a deeper discussion of these levels of legal reasoning is given in [8]. An interesting topic for future work would be to investigate further the relations between the application presented in Chapter 8 and other work dealing with levels of reasoning in legal cases.

As noted in the presentation of the legal example, the particular arguments being modelled in this application did not require any distinction to be made between the consequences that follow from the action (what are referred to as ‘state S’ in argument scheme AS1) and the goal that is a subset of S (what is referred to as ‘goal G’ in argument scheme AS1). S was omitted solely due to the fact that G represents the relevant subset of S in the particular case and no arguments were introduced that required this distinction to be made. If however, such arguments had arisen, then S could easily have been re-instated. Even though such arguments do not arise in the particular example of Chapter 8, it would still be possible to include S in the instantiations of the argument scheme. However, in this case, the statement of S would be the same as that given for the goal G, since S differs from R only by the decision made in A. As legal decisions involve determining what is to be regarded as true, there can be no circumstances, in this example, under which S does not result from performing A in R, and so S cannot come under dispute here. This point reflects the fact that in different contexts different parts of the argument scheme may be of importance, whilst others have less relevance to the particular example or domain.

Looking towards practical realisation of the application in Chapter 8, one major difficulty that would arise would be in representing the extensive knowledge required to model the instrumental reasoning at level one. It is difficult to imagine this kind of knowledge being available in advance of the case. For a particular case, however, it is less difficult to construct the fragment required to drive the reasoning in a particular circumstance. Thus it would be possible to analyse particular cases in this way, as has been done here. Such an analysis provides a useful way of identifying the possible points of contention, the differences in beliefs, desires and values which motivate them, and the level at which the disagreements occur.

9.4 Evaluation of the General Approach

The three example applications that I have used to demonstrate the use of my approach to practical reasoning in BDI agents have each shown interesting features that arise from consideration of arguments in the different domains. Having tested the model in three different contexts I shall now provide an evaluation of the general underlying approach. The criteria against which I shall evaluate the approach are as follows:

- Comprehensiveness of the model.
- Flexibility of the model when used in different domains.
- Realistic representation of real-life arguments.

I have chosen these criteria as they represent three important elements that fit with the aims of this thesis, as stated in Section 1.2 and summarised here: to provide a realistic theory of persuasion in practical reasoning that can be represented for use in agent systems in a number of different domains.

Firstly, I examine the comprehensiveness of the model. The aim of the underlying theory of persuasion over action, given in Chapter 3, was to provide a model of persuasive argument that could capture the nature of all disputes that may occur in matters of practical action. The general argument scheme for practical reasoning explicitly represents the different elements of a justification for action that all need to be accounted for in practical reasoning i.e., the circumstances, the action, the consequences, the goal and the value. In the example applications, no justification for action was encountered that could not be captured by this argument scheme. In order to check that no attack had been omitted from the theory it was my intention to choose example scenarios from domains that would exploit the different types of arguments and attacks from my underlying theory. I hoped to make use of different sets of criticisms to show the importance of having the distinct attacks to unambiguously identify all arguments involved, and to also check that no argument was encountered that could not be represented by an attack from my theory. This has proved to be the case. Each example contained differing sets of attacks that were pertinent to each debate and showed that context plays a large role in the nature of disputes. No arguments were encountered in any of the debates that were not easily captured by one of the attacks from the theory. From this, I am satisfied that underlying theory provides a comprehensive classification of the different arguments that can be posed against a justification for action in the form of my argument scheme.

Secondly, I examine flexibility of the model when used in different domains. I aimed to show that my theory provides a *general* underlying model of persuasive argument in practical reasoning, but this can be used in significantly different domains, without the need for changes to the model. The purpose of the example applications was to show that the general model can handle the differences that are inherent in each domain chosen. This too has proved to be the case because although the examples only represent subsets of the arguments that could be used in each case, no argument/attack was discovered that could not be represented by the model in a natural manner. No arguments were 'forced' to fit the model, even though the debates involved in each example were all of a very different nature. It is important to note that the three domains chosen as settings for the example applications are ones which are inherently focused on actions and decision making. So, it comes as no surprise that models of practical reasoning, such as my own, are suitable for application in these fields.

Thirdly, I examine how realistic my account is in representing real-life arguments. A major aim of my approach was to provide an account of persuasive argument in practical reasoning that could be used by autonomous agents, but at the same time be realistic in its representation of arguments. I believe that my approach meets this target in two ways. Firstly, by making use of *values* I was able to model the different preferences of the parties involved in the debates, which show how different *audiences* have different views on the topics involved. Secondly, I was able to model rational disagreement by showing how disputes concerning values can account for the fact that it is not always possible to reach agreement on the justification of actions. Individual preferences and rational disagreement are both concepts that commonly appear in human reasoning and debates, as discussed in Chapter 2. Representation of these concepts in my theory provides a realistic and useful model in which the true nature of practical reasoning can be captured. Furthermore, I have specified a means by which the model can be represented for use in BDI agents and I believe that this model provides an effective and plausible method that contributes towards the development of realistic practical reasoning in BDI agents. Additionally, I chose scenarios for the applications that represented typical real-life debates involving argumentation, in order to show that the model is applicable to actual scenarios, which has proved to be the case.

Through evaluation against the above criteria I am satisfied that my test applications are able to show the worth and comprehensiveness of my approach. However, I do point out that I made a number of choices, with regard to the inclusion of specific existing approaches, within the proposal of my own theory. For example, I chose to follow Walton's account of practical reasoning which makes use of presumptive arguments in the form of argument schemes and critical questions. Likewise, I chose to make use of Bench-Capon's Value-Based Argumentation Frameworks for the evaluation of arguments. In both these cases there are numerous other similar approaches that I could have made use of. As discussed in Section 2.1, numerous commentaries and models have been given for practical reasoning, and as discussed in Section 2.5.3, extensions and variations to Dung's Argumentation Frameworks have been proposed by a number of different people. In choosing to make use of specific approaches to certain problems it has not been my intention to deny the worth of other similar approaches. The choices were made as a selection of plausible and coherent approaches to the problems. Thus, it may well be the case that other approaches would have proved to be equally useful in meeting my requirements. However, as I have discussed above, the approaches that I have chosen to follow have successfully provided me with the means by which I can demonstrate the worth of my model.

One additional point that should also be noted in evaluating the approach concerns the definition and usage of values. As has been shown in each example, values are highly context dependant elements of arguments and they are personal to the individual and/or group subscribing to the particular values of concern. Hence, the definition and usage of values can be difficult to standardise, though this of course reflects the situation in real life arguments. In the legal debate the values involved relate to social behaviour and the necessity to uphold legal principles, which are themselves formed from social values. In the medical example the values involved relate more to the different perspectives represented by the different information sources in the system. In this way each information agent had their own individual values they were trying to uphold. In the political example the values relate to ethical issues that were associated with the consequences of the proposed actions. Thus, we can see that values vary widely depending upon the domain and the particular situation. It may even be the case that something which can be seen as a value in one particular situation may serve as a goal in another situation. For example, in a medical setting, one value might be 'good health' and this would be promoted by fulfilling the goal of curing a patient of an illness. But, in a wider context a goal may be for patients to have good health in order to promote the value of 'longevity'. This point introduces issues in dealing with the granularity involved in the construction of arguments. Such issues are similar to those involved in deciding upon the granularity of goals to pursue in standard BDI models. For example, an agent may have a goal 'to be happy', and then commit to finding a way to achieve happiness. This may involve adopting a sub-goal of 'being rich', which may itself involve a sub-goal of 'conducting lots of business deals', and so on. The same could be said of values. However, for the purposes of the examples I have given in this thesis I have used values that are relevant to particular contexts and the values recognised by agents would be provided by the designer (as are goals), though individual agents will subscribe to their own particular subset of these. Following the discussion from Section 2.1.1 of values and their importance in human reasoning, I believe that the use of such values in my model reflects the situation in real life where values (and also goals) are specific to individual people and contexts involving decision making. The purpose of the theory proposed in this thesis was to attempt model the subjectivity involved in decision making about actions and give a realistic account of practical reasoning that can be made computational for use in agent systems. I believe that the use of values in my account, although non-standardised and highly context dependant, significantly contributes to this aim, as reflected in the examples presented in the previous three chapters.

9.5 Summary

In this chapter I have individually assessed and discussed each of the applications of my account of practical reasoning in BDI agents, which featured in the preceding three chapters. Following the discussion of the individual examples, I have also given an evaluation of the general underlying approach. Each example has proved to yield its own interesting features and insights and in doing so, the examples have met my objectives regarding the underlying approach.

Additionally, I do also note that there are a number of issues that still need to be addressed in order for my account to be fully effective for deployment in agent systems. These issues have been touched upon in the above evaluations and further issues that remain to be addressed are discussed in more detail in the concluding chapter that follows. However, I do believe that the example applications I have discussed here have proved to be instructive, insightful and successful in achieving their objectives.

Chapter 10

Conclusions and Future Directions

Socrates: "*No greater misfortune could happen to anyone than that of developing a dislike for argument.*"
"*Phaedo*", Plato, Greek Philosopher, (c.427 – c.347 BC).

In this chapter I provide a summary of the contributions made by the work presented in this thesis and I also discuss some areas for possible future work.

10.1 Summary of Contributions

The aim of this thesis, as defined in Chapter 1, was to attempt to answer the following question:

By what means may autonomous software agents make, question, defend and jointly reason about proposals for action?

Throughout the previous chapters I have addressed a number of issues that all contribute towards answering the above question and also address the more specific research goals set out in Section 1.2. Below I summarise my contributions and state how they address my research goals.

In Chapter 3 I articulated a theory of persuasion over action for use in situations involving practical reasoning. The theory contains a number of features that enable it to deal with defeasible reasoning. Firstly, the theory is represented as presumptive reasoning whereby a justification for action can be structured into an argument scheme with associated critical questions. The theory follows the account of Walton and also proposes an extension to his work. This form of argument representation enables the reasoning to be treated in a defeasible manner: presumptions for action are held to be true unless demonstrated otherwise through the use of critical questioning. The persuasive element is dealt with through a comprehensive list of attacks that can be made against the presumption. Such attacks are posed by an opponent in an attempt to persuade the proponent of the action that his presumption does not hold. The list of attacks also reflect the subjective and objective criticisms that can be posed against the presumption. In defining this theory I have addressed the first of my research goals:

> "To provide a theory of persuasion within the setting of practical reasoning which accounts for the defeasible nature of reasoning about action."

Additionally, one of the aims of the theory was to provide a *rational* account of practical reasoning, as stated in my second research goal:

> "Following Searle's account of rationality in action, to separate the objective and subjective components within my theory to provide an explanation of how and why rational disagreement can and does occur in practical reasoning."

I believe that the theory I have provided does indeed meet this aim. The theory draws upon a number of important philosophical observations made by Searle regarding the features of practical reasoning, as discussed in the literature survey in Chapter 2. My theory is intended to overcome some of Searle's objections to the classical model of decision theory, which explain why there cannot be a deductive logic of practical reasoning. Thus, my account is intended to give a more effective and rational approach to practical reasoning in agent systems than accounts that are based upon classical decision theory. My theory addresses these aims through its defeasible nature, and by enabling the distinction to be made between objective and subjective arguments, through explicit use of *values* to represent the personal interests of individuals, in addition to their beliefs and goals. Such values are intended to distinguish between goals that agents want to achieve and the motivating purposes for having these goals. This enables objective arguments to be exchanged which involve the facts of the world, in addition to subjective arguments which involve discussion of personal interests and values. The incorporation of such values means that rational disagreement can be accounted for in the theory. This conforms to Searle's opinion that any rational account of practical reasoning should be able to handle rational disagreement. The theory I presented in Chapter 3 is intended to serve as a comprehensive model by which subjective presumptive justifications for action can be presented and subsequently challenged and defended in a persuasion setting. In defining this theory I have provided the first step towards answering my overall research question. Moving towards the realisation of this theory for use in software agents, my third research goal was:

> "To show how my theory of persuasion over action can be transformed into a computational account that can be effectively deployed in autonomous software agent systems. By using the theory, software agents should be able to recognise the objective and subjective components involved in the reasoning, and respond to criticisms appropriately."

In Chapter 4 I showed how my theory of persuasion over action could be articulated in a more formal representation to form the basis of a dialogue game protocol. The protocol, named PARMA, is primarily defined by a syntax and an axiomatic semantics. PARMA enables instantiations of the argument scheme and critical questions from Chapter 3 to be expressed in a dialogical exchange, whereby participants publicly commit to statements regarding their stance on the matter. The protocol may enable two or more participants to propose, attack and defend proposals for action, in accordance with the underlying theory of Chapter 3. I concluded the discussion of the PARMA Protocol from Chapter 4 with a description of an implementation of a Java program that embodies the protocol, with the program acting as a mediator to enforce compliance with the protocol. This implementation successfully encoded the protocol, though it did raise a number of issues regarding representation and useability of the program, as summarised in Chapter 4. These issues were addressed in the PARMENIDES system, which was discussed in Chapter 6. The protocol and Java implementation provided a useful proof of concept of the theory, allowing human participants to conduct dialogues about action in accordance with the protocol, and it also provided a step along the way to the representation of the theory for use in autonomous software agents.

In Chapter 5 I took the PARMA Protocol forward by showing how the underlying theory of persuasion over action can be made computational for use in BDI agents. I addressed this first by proposing an extension to the BDI agent architecture to include the notion of values, then by showing how it is possible to model a deliberation process for BDI agents. For the first part of this deliberation process, the option generation phase, I gave informal descriptions and a formal specification of pre-conditions to allow BDI agents to put forward and attack presumptive justifications for action, in accordance with my theory. I extended the formal definitions with a proposal to show how time, certainty factors and degrees of promotion can be incorporated into the model. The second part of this deliberation process, the filtering phase, involves an action being chosen as an intention by an agent. This is done through the evaluation of competing arguments by way of an abstract method of argumentation to determine the best action to take, depending on the agents' values. The account I have provided fulfils my goal of articulating my theory of persuasion over action in terms to enable its computational use by BDI agents. The effectiveness of the account was demonstrated in three theoretical applications, and these examples contribute to addressing my final research goal:

“To provide theoretical examples of how my computational account can be used by BDI agents in a number of different domains which involve reasoning about actions. The examples should show that it is important to model subjectivity in practical arguments, and in doing this, my computational account should effectively model human practical reasoning, more so than traditional decision theoretic accounts.”

Chapters 6, 7 and 8 showed how the methods described in Chapter 5 can be applied to example scenarios of practical reasoning in three different domains: politics, medicine and law. Additionally, Chapter 6 provided a description of a human mediation system, called PARMENIDES, based in the domain of eDemocracy. The PARMENIDES system was designed to address the problems of representation and useability that emerged from the Java implementation of the PARMA Protocol, as discussed in Chapter 4. The PARMENIDES system again embodies the theory of persuasion over action but in an entirely different format. The system is intended as an online mediation tool whereby a presumptive justification for action, based on a real-life political debate, is presented to the user for his consideration. The user is then led, in a structured fashion, through a series of web pages which decompose the justification into elements that the user is able to question and disagree with. He is then given the chance to input his views on the matter and this is all done in accordance with the underlying theory. PARMENIDES successfully overcomes many of the problems identified with the Java program, as it requires no prior knowledge of the underlying theory for its use, and it could easily be tailored to represent a wide variety of issues involving practical reasoning.

The same topic of discussion as used in the PARMENIDES system was then put to use in a second eDemocracy application in Chapter 6, but this time the application was intended solely for use by BDI agents. The intention of this example was to demonstrate that BDI agents can effectively reason about real-life political issues, in terms of my account. Typically, such issues involve a number of opinions that are based upon personal interests and values and this example was intended to show how my account can be successfully used in a domain that often involves subjective arguments. A discussion of the relation between this application and the PARMENIDES system was also given in Chapter 9, as both applications make use of the same political issue.

The medical application of Chapter 7 described how reasoning about the treatment of a patient can be conducted by a value enhanced BDI agent. Contrary to the eDemocracy example from Chapter 6, only one agent was involved in the decision making and this agent gathered data from a number of specific sources of information in order to perform its reasoning.

Finally, the legal application of Chapter 8 showed how a well-known case from property law can be represented in terms of my model. This example was intended to show how a number of agents with independent knowledge bases can contribute their opinions to legal decision making involving two adversaries.

These three BDI agent applications have all been successful in meeting their aims and they have provided proof that arguments over action can be specified in terms of BDI agents augmented to deal with values, in a wide range of domains. A summary and evaluation of the applications was presented in Chapter 9 along with a discussion of the individual features of each application.

The results and investigations presented in this thesis draw upon a number of different areas in an attempt to answer my research question. The contributions I have made in addressing these issues have produced a comprehensive theory of persuasion in practical reasoning and have shown how this can be formally represented for use by BDI agents in a number of domains.

10.2 Future Directions

The results presented in this thesis have provided a number of possible directions for future work. Firstly, an interesting avenue to pursue would be an analysis of strategies for successful persuasion. It may be the case that certain types of attack from the theory of persuasion are more effective than others in having a persuasive effect on an opponent. For example, it may be easier to persuade an agent to believe that the effects of an action are not as first thought, rather than persuade the agent that achievement of a desire does not promote a particular value. Furthermore, some attacks may prove to be more persuasive in one particular domain than in another. In order to make an assessment of the persuasive power of the individual attacks, real-life applications involving implemented agents arguing over a wide range of practical topics would be required, and this is itself an obvious area for future work. Such an implementation and analysis would be a large task and is thus outside the scope of this thesis.

A second obvious area to address in future work would be the relation between the account of persuasion given here and other theories of persuasive argument specified for different settings of agent interaction. For example, accounts of negotiation based upon the exchange of arguments for persuasive effects, as discussed in Chapter 2. There may be connections and insights that can be drawn between the methods and motivations presented in these and other similar accounts of persuasive argument and the model presented in this thesis.

A third factor that needs to be investigated in order for my account to deployed in agent systems is scalability. I am satisfied that each of the example applications detailed in Chapters 6, 7 and 8 contain a sufficient range of varying arguments to show the effectiveness of the approach. However, in domains where decisions about actions are highly complicated further investigation needs to be conducted to examine how tractable my account is to large scale complex domains. I believe that the theory of practical reasoning which my account is grounded upon is comprehensive as an underlying model of argument. However, I also believe that in order to address such issues of scalability the theory of persuasion would need to be supplemented by heuristics to ensure it could be used efficiently in all situations. For example, consideration needs to be given to the granularity of the decision making. It may be the case that reasoning may proceed on different levels of scope whereby the reasoning at one level may be required to be completed before reasoning at a higher level can be considered. Such a situation was involved in the legal example of Chapter 8. Here the bottom level reasoning concerned facts about the world that needed to be decided upon before reasoning could proceed at the next level which linked facts about the world to legal concepts. In such a situation agents need to be equipped with the ability to recognise the order in which decisions must be made. One way of tackling this would be to provide strategies or heuristics to guide the agents in decomposing decision problems. This would enable agents to address decisions in order of their importance to enable the issue in question to be resolved. The decomposition of decision problems is also related to the concept of granularity of plans. In my model it is assumed that once an agent has decided upon an intention to commit to, it retrieves a plan from its plan library to realise this intention and thus the granularity of the plan is pre-determined. Tied in with this is the consideration that needs to be given to resource bounded agents. The factors that circumscribe the critical questions that need to be posed could encompass a number of things including: the agent's beliefs and desires, limits on the time afforded to the agent to make a decision, the consequences of belief revision, and so on. One particularly notable discussion of issues related to planning in AI has been given by Pollack in [125]. However, investigation of these factors falls outside the scope of this thesis: the account I have presented is intended as a foundational model to capture the important features of practical reasoning that were discussed in Chapter 2. Nonetheless, I do believe that an instructive area for future investigation would be to examine extensions to this foundational model to enable it to be scaled up to deal with large and complex decision making tasks.

A fourth area to address in future work would be to extend elements of my account that do not currently form the primary focus, but would provide useful extensions to it, were they to be developed further. The first of these elements is the notion of accrual, as discussed in the eDemocracy example of Chapter 6. The possibility of capturing such a notion in my account emerged from the debate involved in the example in Chapter 6 and it was not one of my original intentions to model this notion. However, I believe that my account is able to give some insight into how accrual of arguments can be represented and I proposed an outline of how this could be dealt with in Section 6.5. Further investigation and articulation of this element would be an interesting area for future work. A second area for extension within my account concerns a point that was outlined in Chapter 5 regarding the use of time, certainty factors and degrees of promotion. In Section 5.4 I provided an outline of how the definitions for the BDI application of my theory can be extended to allow time, uncertainty and degrees of promotion to be dealt with. It is often the case that agents will operate in environments where information regarding the current state of the world, the viability of actions, the outcome of actions or the degree to which goals promote values will be uncertain. The focus of my account has been to provide a comprehensive underlying theory to enable autonomous software agents to rationally reason about decisions over action. So, my immediate focus has not concerned issues of time and uncertainty. However, the proposal I have provided in Section 5.4 demonstrates how my account can in fact include such elements and I provided a small example to demonstrate the use of such features. Further investigation and articulation of such elements within my account would be a useful extension for future work and is something that could be explored in more comprehensive examples.

Finally, I have said little here about the social effects of relationships between agents situated in a multi-agent system, and there are a number of interesting issues to note about this. Individual agents are often party to one or more particular groups with their own beliefs, desires and values. Thus, individuals in a group form opinions and have desires and values based upon the norms and interests of the group. Social interests not only affect an individual's position on particular issues, but they can also affect the likelihood of persuasion being successful or unsuccessful in a particular matter. For example, persuasion is much more likely to occur in scenarios where decisions do not concern matters based upon individuals' personal ethics and values. It may also be the case that hierarchies exist between certain roles that affect the power of persuasion. For example, in a parent-child relationship the parent usually has the final say on practical matters that the child is involved in and thus would not be easily persuaded to change their mind on issues of importance. There is a large amount of research being conducted into the social aspects of multi-agent systems (e.g., [44, 52, 118, 151]) and one interesting direction for future investigations would be to explore the relationship between social interactions and the effects of persuasion in practical reasoning in different social scenarios.

The areas that I have indicated here for possible future directions are just some of the options presented by the work detailed in this thesis. There are also a number of interesting sub-issues that would benefit from further investigation. The results presented here are intended only to provide details of how persuasion in practical reasoning can be dealt with in BDI agents. For agents to be fully autonomous in dealing with practical reasoning the theory presented here also needs to be complemented by other aspects of agency, such as: reasoning about beliefs and belief revision, successful completion and evaluation of actions and their effects, strategies for maximising persuasion, fully integrated communication capabilities, and methods for planning and plan revision, to name but a few areas. Nonetheless, I believe that the findings reported in this thesis provide a background by which one particular area of philosophy that is central to multi-agent systems – practical reasoning – can be built upon in the quest for the effective design and construction of realistic automated computer systems.

Appendix A

A Denotational Semantics of the PARMA Protocol

In this appendix I present an outline of a denotational semantics for the PARMA Protocol, presented in Chapter 4. These semantics were developed in joint work with Peter McBurney and Trevor Bench-Capon and have been published in [10] and [15].

A.1 Definitions

As discussed in Section 2.6.2, a denotational semantics is a semantics which maps statements in the syntax of a language to mathematical entities [156]. The approach taken here draws on a branch of category theory, namely topos theory. Our reason for using this, rather than (say) a Kripkean possible worlds framework or a labelled transition system, is that topos theory enables a natural representation of logical consequence ($S \models G$) in the same formalism as mappings between spaces ($R \overset{A}{\to} S$ and $G \uparrow v$). To our knowledge, no other non-categorical denotational semantics currently proposed for action formalisms permits this.

We begin by representing proposals for action. We assume, as in Section 3.2, finite sets of Acts, Propositions, States, Goals, and Values, and various mappings. For simplicity, we assume there are n propositions. Each State may be considered as being equivalent to the set of propositions which are true in that State, and so there are 2^n States. We consider the space $\mathcal{C}$ of these States, with some additional structure to enable the representation of actions and truth-values. We consider elements of Values to be mappings from Goals to some space of evaluations, called $\mathcal{S}$. This need not be the three-valued set *Sign* $= \{+, =, -\}$ that was assumed in Section 3.2, although we assume that $\mathcal{S}$ admits at least one partial order. The structures we assume on $\mathcal{C}$, on $\mathcal{S}$ and between them are intended to enable us to demonstrate that these are categorical entities [69]. We begin by listing the mathematical entities, along with informal definitions.

- The space $\mathcal{C}$ comprises a finite collection $\mathcal{C}_0$ of objects and a finite collection $\mathcal{C}_1$ of arrows between objects.

- $\mathcal{C}_0$ includes 2^n objects, each of which may be considered as representing a State. We denote these objects by the lower-case Greek letters, $\alpha, \beta, \gamma, \ldots$, and refer to them collectively as *state objects* or *states*. We may consider each state to be equivalent (in some sense) to the set of propositions which are true in the state.

- $\mathcal{C}_1$ includes arrows between state objects, denoted by lower case Roman letters, $f, g, h, \ldots$. If f is an arrow from object α to object β, we also write $f : \alpha \rightarrow \beta$. Some arrows between the state objects may be considered as representing actions leading from one state to another, while other arrows are causal processes (not actions of the dialogue participants) which take the world from one state to another. There may be any number of arrows between the same two objects: zero, one, or more than one.

- Associated with every object $\alpha \in \mathcal{C}_0$, there is an arrow $1_\alpha \in \mathcal{C}_1$ from α to α, called the identity at α. In the case where α is a state object, this arrow may be considered as that action (or possibly inaction) which preserves the status quo at a state α.

- If $f : \alpha \rightarrow \beta$ and $g : \beta \rightarrow \gamma$ are both arrows in $\mathcal{C}_1$, then we assume there is an arrow $h : \alpha \rightarrow \gamma$. We denote this arrow h by $g \circ f$ (*"g composed with f")*. In other words, actions and causal processes may be concatenated.

- We assume that $\mathcal{C}_0$ includes a special object *Prop*, which represents the finite set of all propositions. We further assume that for every object $\alpha \in \mathcal{C}_0$ there is a monic arrow $f_\alpha : \alpha \rightarrow Prop$. Essentially, a monic arrow is an injective (one-to-one) mapping.

- We assume that $\mathcal{C}_0$ has a terminal object, **1**, i.e., an object such that for every object $\alpha \in \mathcal{C}_0$, there is precisely one arrow $\alpha \rightarrow \mathbf{1}$.

- We assume that $\mathcal{C}$ has a special object Ω, and an arrow $true : \mathbf{1} \rightarrow \Omega$, called a *sub-object classifier*. The object Ω may be understood as the set comprising $\{True, False\}$.

- We assume that $\mathcal{S}$ is a space of objects over which there is a partial order $<_i$ corresponding to each participant in the dialogue. Such a space may be viewed as a category, with an arrow between two objects α and β whenever $\alpha <_i \beta$. For each participant, we further assume the existence of one or more mappings v

between $\mathcal{C}$ and $\mathcal{S}$, which take objects to objects, and arrows to arrows. We denote the collection of all these mappings by $\mathcal{V}$.

The assumptions we have made here enable us to show that $\mathcal{C}$ is a category [69], and we can thus represent the statement $R \xrightarrow{A} S$, for states R and S, and action A. Moreover, the presence of a sub-object classifier structure enables us to represent statements of the form $S \models G$, for state S and goal G, inside the same category $\mathcal{C}$. This structure we have defined for $\mathcal{C}$ creates some of the properties needed for $\mathcal{C}$ to be a topos [69]. Finally, each space $\mathcal{S}$ with partial order $<_i$ is also a category, and the mappings v are functors (structure-preserving mappings) between $\mathcal{C}$ and $\mathcal{S}$. This then permits us to represent statements of the form $G \uparrow v$, for goal G and value v.

We define a denotational semantics for the PARMA Protocol by associating dialogues conducted according to the protocol with mathematical structures of the type defined above. Thus, the statement of a proposal for action by a participant in a dialogue

$$R \xrightarrow{A} S \models G \uparrow v$$

is understood semantically as the assertion of the existence of objects representing R and S in $\mathcal{C}$, the existence of an arrow representing A between them, the existence of an arrow with certain properties[1] between *Prop* and Ω, and the existence of a functor $v \in \mathcal{V}$ from $\mathcal{C}$ to $\mathcal{S}$. Attacks on this position then may be understood semantically as denials of the existence of one or more of these elements, and possibly also, if the attack is sufficiently strong, the assertion of the existence of other objects, arrows or functors.

Thus, our denotational semantics for a dialogue conducted according to the PARMA Protocol is defined as a countable sequence of triples,

$$\langle \mathcal{C}_1,\ \mathcal{S}_1,\ \mathcal{V}_1 \rangle,\ \langle \mathcal{C}_2,\ \mathcal{S}_2,\ \mathcal{V}_2 \rangle,\ \langle \mathcal{C}_3,\ \mathcal{S}_3,\ \mathcal{V}_3 \rangle,\ \ldots,$$

where the k-th triple is created from the k-th utterance in the dialogue according to the representation rules just described. Then, our denotational semantics for the PARMA Protocol itself is defined as the collection of all such countable sequences of triples for valid dialogues conducted under PARMA. This approach views the semantics of the protocol as a space of mathematical objects, which are created incrementally and jointly by the participants in the course of their dialogue together. The approach derives from the constructive view of human language semantics of Discourse Representation Theory [89], and is similar in spirit to the denotational semantics, called a *trace semantics*, defined for deliberation dialogues in [108], and the *dialectical graph* recording the statements of the participants in the "Pleadings Game" of Gordon [70]. In future work

[1] This arrow is the characteristic function for the object representing G, and the properties are that a certain diagram commutes in $\mathcal{C}$.

we hope to extend the outline of the denotational semantics given here by further extending the definitions to enable us to study and reason about properties of the PARMA Protocol.

Appendix B

Design Documentation for the Java Dialogue Game

In this appendix I present the design documentation for the implementation of the Java dialogue game that is based upon the PARMA Protocol and is described in Section 4.4. Section B.1 presents the analysis and design documentation for the implementation. Section B.2 describes the implementation and testing. Section B.3 provides a brief discussion of the implemented system.

B.1 Analysis and Design

Firstly I present an analysis of the Java classes that are needed to encode the solution, then I give detailed design tables for the implementation.

Figure B.1 presents a primitive class diagram showing the main classes that are needed for the dialogue game implementation. The code actually makes use of many more pre-defined classes from the Java Applications Programming Interface (API) but they have been omitted from this design documentation because although they are necessary for the program to function correctly, they are not the main focus point of the implementation presented here. The classes are all represented in the form of simplified UML style diagrams.

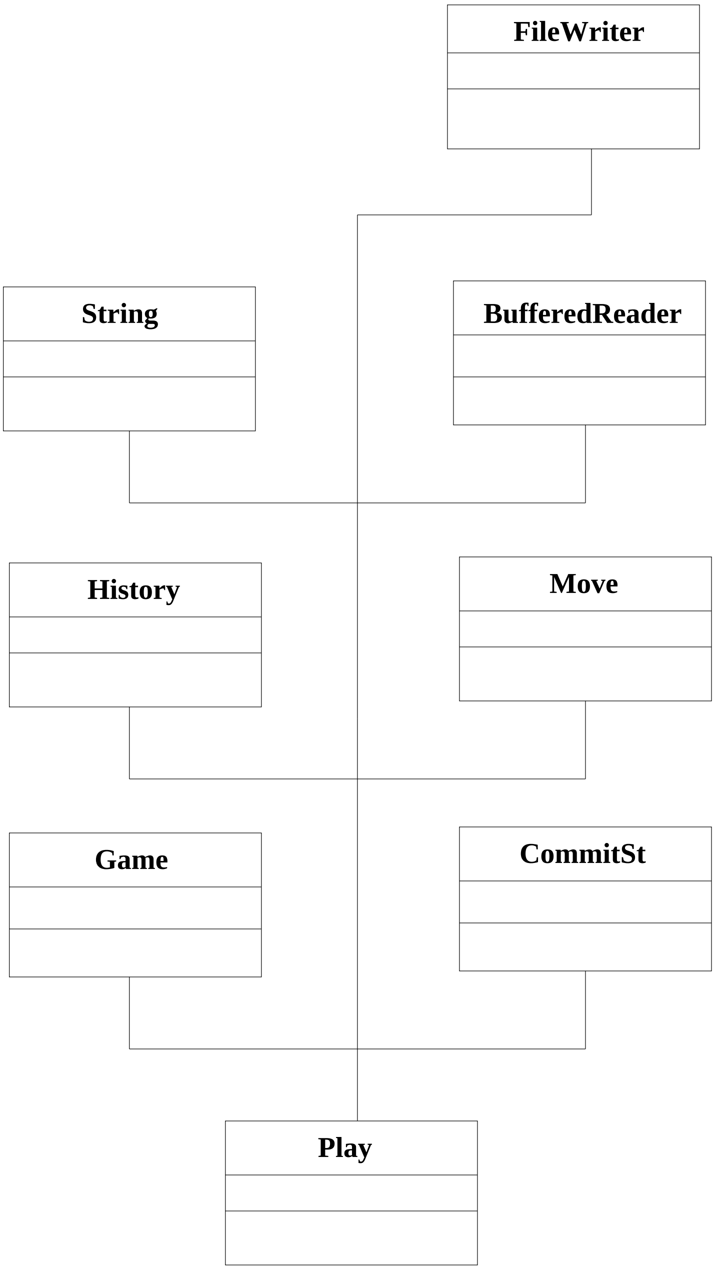


Figure B.1: Primitive class diagram.

Given below in tables B.1 – B.9 is an analysis for each of the individual classes shown in Figure B.1. Each of the UML style diagrams representing a class shows the fields, constructors and methods that are used in that individual class.

Table B.1: **History Class**

History
private String[][] history
public History() public void updateHistory(String, String) public void legalUpdate(String) public void illegalUpdate(String) public void printHistory()

Table B.2: **Move Class**

Move
private boolean movePossible private String[] possibleMoves private static final int LOWER_BOUND private static final int ARRAY_LENGTH
public Move() public void checkPossible(String, String, String, String, String, CommitSt, Game, History, Move) public void successful(String, String, String, String, String, CommitSt, Game, History, Move)

Table B.3: **Game Class**

Game
private String speaker private String hearer private String[] possibleMoves public static BufferedReader keyboardInput
public Game() public void firstMove(String, String, CommitSt, Game, History, Move) public void turnFinished(String, String, String, String, String, CommitSt, Game, History, Move) public void makeMove(String, String, String, String, String, CommitSt, Game, History, Move)

Table B.4: **CommitSt Class**

CommitSt
private String[][] play1ComSt private String[][] play2ComSt private String content public static BufferedReader keyboardInput
public CommitSt() public void answerAsk(String, String, String, String, String, CommitSt, Game, History, String, String, Move) public void askAccept(String, String, String, String, String, CommitSt, Game, History, String, Move) public void illegalMove(String, String, String, String, CommitSt, Game, History, String, History, String, Move) public void legalAccept(String, String, String, String, String, CommitSt, Game, History, String, Move) public void legalStateCirc(String, String, String, String, String, CommitSt, Game, History, Move) public void legalStateAction(String, String, String, String, String, CommitSt, Game, History, Move) public void legalStateConseq(String, String, String, String, String, CommitSt, Game, History, Move) public void legalStateLogCons(String, String, String, String, String, CommitSt, Game, History, Move) public void legalStatePurp(String, String, String, String, String, CommitSt, Game, History, Move) public void legalDenyCirc(String, String, String, String, String, CommitSt, Game, History, Move) public void legalDenyConseq(String, String, String, String, String, CommitSt, Game, History, Move)

Table B.5: **CommitSt Class continued**

CommitSt
public void legalDenyLogCons(String, String, String, String, String, CommitSt, Game, History, Move) public void legalDenyPurp(String, String, String, String, String, CommitSt, Game, History, Move) public void legalDenyInitCircExist(String, String, String, String, String, CommitSt, Game, History, Move) public void legalDenyActExist(String, String, String, String, String, CommitSt, Game, History, Move) public void legalDenyNewStateExist(String, String, String, String, String, CommitSt, Game, History, Move) public void legalDenyGoalExist(String, String, String, String, String, CommitSt, Game, History, Move) public void legalDenyValueExist(String, String, String, String, String, CommitSt, Game, History, Move) public void legalAskCirc(String, String, String, String, String, CommitSt, Game, History, Move) public void legalAskAct(String, String, String, String, String, CommitSt, Game, History, Move) public void legalAskConseq(String, String, String, String, String, CommitSt, Game, History, Move) public void legalAskLogCons(String, String, String, String, String, CommitSt, Game, History, Move) public void legalAskPur(String, String, String, String, String, CommitSt, Game, History, Move) public String p1CheckDenial(String, String, String, String) public String p2CheckDenial(String, String, String, String) public void printComStores(String, String)

Table B.6: **Play Class**

Play
public static BufferedReader keyboardInput
public Play() public static void main(String[] args)

Table B.7: **String Class**

String
public boolean equals(String) public int compareTo(String) public String substring(int,int) public int indexOf(int) public String concat(String)

Table B.8: **BufferedReader Class**

BufferedReader
public String readLine()

Table B.9: **FileWriter Class**

FileWriter
public void close() public void write(String)

The tables presented above define the fields and methods used in each class of the Java implementation of the PARMA Protocol. Additionally, I have specified each of the above classes in further detail in summary tables explaining the purpose and interaction of all the fields, constructors and methods in each class. The methodology used to construct these tables follows the format used by Sun to describe the Java Applications Programming Interface (API), which can be found at: `http://www.java.sun.com/reference/api/index.html`. The design tables I have specified for the PARMA Protocol have been omitted here for reasons of space, though they can all be found in [11].

Furthermore, the state transition diagram for the dialogue game, presented in Section 4.4, is intended as a high level supplement to the design documentation presented in this appendix.

B.2 Implementation and Testing

The dialogue game deign given in the previous section was implemented in the Java programming language and it strictly follows the design given in the above section. The author can be contacted if more details of the actual implementation are needed.

On completion of the implementation the program was tested to ensure that all the locutions specified in the axiomatic semantics for the PARMA Protocol, given in Section 4.2, could be executed as expected. The testing strategy simply followed the verification of moves being successfully executed in accordance with the axiomatic semantics. Each set of locutions was individually tested to check that all pre-conditions needed for the execution of a move were met and that all post-conditions were correctly applied. The testing has proven this to be the case and there are no known bugs. As the test cases were all formulated from the conditions set out in the axiomatic semantics I have not included the test data here and I refer the reader back to Section 4.2 for this information. I now provide an example transcript showing the program in use and this will be followed by a brief evaluation of the implemented program.

B.2.1 Example Use of the Dialogue Game

Tables B.10 and B.11 give an example transcript of a dialogue game being conducted in accordance with the PARMA Protocol. Two parties, A and B, are engaged in a discussion about where and what type of holiday to go on together. Persuasion occurs through use of the attacks from the protocol, as both parties have different preferences. The dispute arises due to party A wanting to go on a beach holiday and party B wanting to go on a skiing holiday.

Table B.10: **Example Dialogue**

Move No.	Player	Locutions	Content
1	A	state circs(R_1)	I would like to book our summer holiday today.
2	A	ask circs(R)	Would you like to book our holiday today?
3	B	state circs(R_1)	Yes, I would like to book our holiday today.
4	B	ask action(A)	So, where do you think we should go?
5	A	state action(A_1)	Well, I was thinking about one of the Greek islands.
6	A	ask action(A)	Where did you have in mind?
7	B	state action(A_2)	I fancied a skiing holiday somewhere
8	B	state conseq(A_2,R_1,S_1)	so we'd get to go to a ski resort
9	B	state log conseq(S_1,G_1)	where there would be lots of things to do during the day as well as at night
10	B	state purpose(G_1,V_1,D+)	so I think we'd have a really good time.
11	A	*attack 6*: state action(A_1) state conseq(A_1,R_1,S_2) state log conseq(S_2,G_1)	 But going to the Greek islands means we'd get to go to a beach resort where there's also lots to do during the day and at night too.
12	A	state conseq(A_1,R_1,S_2)	And, the Greek islands are nice and hot during the summer months
13	A	state log conseq(S_2,G_2)	and you know how I like to spend my holidays in the sun
14	A	state purpose(G_2,V_2,D+)	as it helps me to relax.
15	B	*attack 9*: state conseq(A_1,R_1,S_3) state log conseq(S_3,G_3) state purpose(G_3,V_3,D-)	 But we went on a similar holiday last year and it'll be just the same and I'd like to do something different this year.
16	B	*attack 9*: state conseq(A_2,R_1,S_4) state log conseq(S_4,G_4) state purpose(G_4,V_2,D-)	 But going skiing means going on holiday to somewhere that's in a cold climate and I want a holiday in the sun.
17	B	*attack 9*: state conseq(A_2,R_1,S_5) state log conseq(S_5,G_5) state purpose(G_5,V_4,D+)	 But it will be an activity holiday which is different from what we're used to and that will make it more exciting.

Table B.11: **Example Dialogue cont'd**

18	B	*attack 6*: state action(A_3) state conseq(A_3,R_1,S_6) state log conseq(S_6,G_2)	And, we can go on a beach holiday in winter instead of our usual city break so we'll get to go on two holidays one of which will be your beach holiday.
19	A	state purpose(G_2,V_2,D+)	I suppose those two holidays would be nice.
20	A	*attack 9*: state conseq(A_2,R_1,S_2) state log conseq(S_2,G_6) state purpose(G_6,V_5,D-)	But going skiing means going on holiday to somewhere that's very expensive and we can't afford to spend a lot of money.
21	B	*attack 7a*: state action(A_4) state conseq(A_4,R_1,S_4) state log conseq(S_4,G_7) state purpose(G_7,V_5,D+)	Not if you go to somewhere in Eastern Europe as there are ski resorts there which are very cheap at the moment and it really won't cost a lot of money.
22	A	state purpose(G_7,V_5,D+)	Yes, that's true. A lot of friends have told me its cheap to go skiing in Eastern Europe.
23	B	ask action(A)	Great. So you agree to a skiing holiday then?
24	A	state action(A_2)	Yes, I'll agree to it.
25	A	state action(A_5)	We can go to the travel agents tomorrow and look at some destinations and prices.

The above transcript represents an example of a dialogue being conducted between two parties but it does not show any details of the internal records made by the program concerning the commitments incurred by the players. The program mediates the dialogue exchange by checking that the pre-conditions for chosen moves hold. If they do hold, the program then executes the post-conditions of the move by updating the appropriate player's commitment store to include the addition of any newly incurred commitments.

The program also maintains a history of all moves chosen throughout the course of the dialogue, even when the move chosen is an illegal one. The history records the details of the player who is making the move, the name of their chosen locution, the status of the locution, which can either be 1 to show that the chosen move is a legal move or -1 to denote that an illegal move was chosen and has not incurred any new commitment, and finally, the history contains the content of the locution chosen. Thus, the history documents all moves attempted and made.

The commitment stores of each player contain the locution names of the moves they have made, the content of each move and the status of each commitment, with 1 being positive commitment to the content, -1 being commitment to the negation of the content and 0 being no commitment as to the truth or falsity of the content.

Both the commitment stores and the history are updated and displayed on screen whenever a new move/commitment is added to either. Tables B.12–B.14 below give example snapshots of the history and each player's commitment store, displaying their contents after moves 1-7 from the example dialogue above have been executed[1]:

Table B.12: **History after move 7**

Player	Move	Status	Content
A	state circs	1	I would like to book our summer holiday today.
A	ask circs	1	Would you like to book our holiday today
B	state circs	1	Yes, I would like to book our holiday today.
B	ask action	1	So, where do you think we should go?
A	state action	1	Well, I was thinking about one of the Greek islands.
A	ask action	1	Where did you have in mind?
B	state action	1	I fancied a skiing holiday somewhere.

Table B.13: **Player A's commitment store after move 7**

Move	Status	Content
enter dialogue	1	enter dialogue
state circs	1	I would like to book our summer holiday today
circ exist	1	I would like to book our summer holiday today exists in set of possible circs
state action	1	Well, I was thinking about one of the Greek islands
act exist	1	Well, I was thinking about one of the Greek islands exists in the set of possible actions

Table B.14: **Player B's commitment store after move 7**

Move	Status	Content
enter dialogue	1	enter dialogue
state circs	1	Yes, I would like to book our holiday today
circ exist	1	Yes, I would like to book our holiday today exists in set of possible circs
state action	1	I fancied a skiing holiday somewhere
act exist	1	I fancied a skiing holiday somewhere exists in the set of possible actions

[1] Note: as a natural language dialogue is being modelled in this example the commitment stores include words that are conversation fillers, such as 'well', 'so', etc, as these are naturally used in everyday conversation. However, if the game were to be used by computer agents, rather than human agents, such words would not be included. In such a case the content of moves would contain purely propositional statements based on the representation of the knowledge embodied in the computer agents' knowledge bases.

B.3 Evaluation

The completed implementation allows two human players to play the game, in accordance with the given specification. However, a number of issues came to light through implementing the dialogue game. A summary of these issues was given in Section 4.4.3, though a detailed evaluation of the implemented system can be found in [11]. Some of the issues encountered in this implementation were addressed and resolved by the PARMENIDES system, which was discussed in Chapter 6.

Although the program has been fully tested and contains no known errors, there are a number of issues that could be addressed in future work. Firstly, no consideration has been given to the efficiency of the code and it may well be that there are more computationally efficient encodings of the protocol. Secondly, the version of the protocol that has been implemented is just one of a number of different versions that the protocol can embody. The version described in this implementation is based upon quite strict pre-conditions for the performance of the locutions. It would however, be possible to design and implement a 'looser' version of the protocol, where the number of pre-conditions to be fulfilled would be reduced. This would increase the flexibility afforded by the protocol, but it may exacerbate the problems relating to the correct use of the protocol, as discussed in Section 4.4.3. As the purpose of this implementation was to provide a proof of concept for the PARMA Protocol, such alternative implementations have not been explored in this body of work. The final point to note in this discussion is that the program is not very user friendly as there is no graphical user interface. It would be possible to construct a simple interface to the program however, the implementation that superseded the Java program, the PARMENIDES system, does provide a user friendly interface so I am satisfied that this issue has been overcome. This concludes my discussion of the Java implementation of the PARMA Protocol.

Appendix C

Abstract Argumentation Frameworks: Definitions

In this appendix I present the definitions for Dung's abstract argumentation frameworks [55] and also the definitions for Bench-Capon's Value-Based Argumentation Frameworks [26]. Both frameworks were discussed in Chapters 2 and 5, and Bench-Capon's system is used in the example applications of Chapters 6, 7 and 8. Section C.1 gives the definitions for Dung's argumentation frameworks and Section C.2 gives the definitions for Bench-Capon's Value-Based Argumentation Frameworks. The definitions presented here are given for reference purposes and they are taken from a paper by Dunne and Bench-Capon [57], to whom I am most grateful for their consent to reproduce the definitions here.

C.1 Basic Definitions for an Argumentation Framework

The basic definition for an Argument System (or Argument Framework as they are also analogously referred to) is given below and is derived from that given in [55].

Definition 1 *An* argument system *is a pair $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$, in which $\mathcal{X}$ is a finite set of* arguments *and $\mathcal{A} \subset \mathcal{X} \times \mathcal{X}$ is the* attack relationship *for $\mathcal{H}$. A pair $\langle x, y \rangle \in \mathcal{A}$ is referred to as 'y* is attacked by *x' or 'x* attacks *y'. For R, S subsets of arguments in the system $\mathcal{H}(\langle \mathcal{X}, \mathcal{A} \rangle)$, we say that*

a. $s \in S$ *is* attacked *by R if there is some $r \in R$ such that $\langle r, s \rangle \in \mathcal{A}$.*

b. $x \in \mathcal{X}$ *is* acceptable with respect to *S if for every $y \in \mathcal{X}$ that attacks x there is some $z \in S$ that attacks y.*

c. S *is* conflict-free *if no argument in S is attacked by any other argument in S.*

d. *A conflict-free set S is* admissible *if every argument in S is acceptable with respect to S.*

e. *S is a* preferred extension *if it is a maximal (with respect to $\subseteq$) admissible set.*

f. *S is a* stable extension *if S is conflict free and every argument $y \notin S$ is attacked by S.*

g. *$\mathcal{H}$ is* coherent *if every preferred extension in $\mathcal{H}$ is also a stable extension.*

An argument x is credulously accepted *if there is* some *preferred extension containing it; x is* sceptically accepted *if it is a member of* every *preferred extension.*

C.2 Basic Definitions for a Value-Based Argumentation Framework

The formal definition of a *value-based argumentation framework* (VAF), as given in [57], is presented below.

Definition 2 *A* value-based argumentation framework *(VAF), is defined by a triple $\langle \mathcal{H}(\mathcal{X}, \mathcal{A}), \mathcal{V}, \eta \rangle$, where $\mathcal{H}(\mathcal{X}, \mathcal{A})$ is an argument system, $\mathcal{V} = \{v_1, v_2, \ldots, v_k\}$ a set of k* values, *and $\eta : \mathcal{X} \rightarrow \mathcal{V}$ a mapping that associates a value $\eta(x) \in \mathcal{V}$ with each argument $x \in \mathcal{X}$. An* audience, *α, for a VAF $\langle \mathcal{H}, \mathcal{V}, \eta \rangle$, is a total ordering of the values $\mathcal{V}$. We say that v_i* is preferred to *v_j* in the audience *α, denoted $v_i \succ_\alpha v_j$, if v_i is ranked higher than v_j in the total ordering defined by α.*

Using VAFs, ideas analogous to those of admissible argument in standard argument systems are defined in the following way. Note that all these notions are now relative to some audience.

Definition 3 *Let $\langle \mathcal{H}(\mathcal{X}, \mathcal{A}), \mathcal{V}, \eta \rangle$ be a VAF and α an audience.*

a. *For arguments x, y in $\mathcal{X}$, x is a* successful attack *on y (or x* defeats *y)* with respect to the audience *α if: $\langle x, y \rangle \in \mathcal{A}$* and *it is* not *the case that $\eta(y) \succ_\alpha \eta(x)$.*

b. *An argument x is* acceptable to the subset *S with respect to an audience α if: for every $y \in \mathcal{X}$ that* successfully attacks *x with respect to α, there is some $z \in S$ that successfully attacks y with respect to α.*

c. *A subset R of $\mathcal{X}$ is* conflict-free with respect to the audience *α if: for each $\langle x, y \rangle \in R \times R$, either $\langle x, y \rangle \notin \mathcal{A}$ or $\eta(y) \succ_\alpha \eta(x)$.*

d. *A subset R of $\mathcal{X}$ is* admissible *with respect to the audience α if: R is conflict free with respect to α and every $x \in R$ is acceptable to R with respect to α.*

e. *A subset R is a* preferred extension *for the audience α if it is a maximal admissible set with respect to α.*

f. *A subset R is a* stable extension *for the audience α if R is admissible with respect to α and for all $y \notin R$ there is some $x \in R$ which successfully attacks y.*

A standard consistency requirement which is assumed of the VAFs considered is that every directed cycle of arguments in these contains *at least two* differently valued arguments. The authors of [57] do not believe that this condition is overly restricting, since the existence of such cycles in VAFs can be seen as indicating a flaw in the formulation of the framework. While in standard argumentation frameworks cycles arise naturally, especially when dealing with uncertain or incomplete information, in VAFs odd length cycles in a single value represent paradoxes and even length cycles in a single value can be reduced to a self-defeating argument. Given the absence of cycles in a single value the following important property of VAFs and audiences was demonstrated in [25].

For every audience, α, $\langle \mathcal{H}(\langle \mathcal{X}, \mathcal{A} \rangle), \mathcal{V}, \eta \rangle$ has a unique non-empty preferred extension, $P(\mathcal{H}, \eta, \alpha)$ which can be constructed by an algorithm that takes $O(|\mathcal{X}| + |\mathcal{A}|)$ steps. Furthermore $P(\mathcal{H}, \eta, \alpha)$ is a stable extension with respect to α.

From the above it follows that, when attention is focused on one specific audience, the decision questions become much easier. There are, however, a number of new issues that arise in the value-based framework from the fact that the relative ordering of different values promoted by distinct audiences results in arguments falling into one of three categories.

C1. Arguments, x, that are in the preferred extension $P(\mathcal{H}, \eta, \alpha)$ for some audiences but not all. Such arguments being called *subjectively acceptable*.

C2. Arguments, x, that are in the preferred extension $P(\mathcal{H}, \eta, \alpha)$ for *every* audience. Such arguments being called *objectively acceptable*.

C3. Arguments, x, that do not belong to the preferred extension $P(\mathcal{H}, \eta, \alpha)$ for *any* choice of audience. Such arguments being called *indefensible*.

Additionally, in [26] Bench-Capon gives a discussion of how uncertainty regarding facts can be dealt with in such value-based systems, as arguments can turn on the support given to facts as well as values. The solution here is to treat unknown facts as though they are values but, in order to give such facts their due weight, they *must always be given the value with the highest preference for all parties*. The addition of this rule introduces four possibilities for the status of arguments:

- objectively acceptable *sceptically*: when the argument appears in every preferred extension,
- objectively acceptable *credulously*: when the argument appears in every preferred extension corresponding to some choice of facts,
- subjectively acceptable *sceptically*: when the argument appears in every preferred extension relating to some value order,
- subjectively acceptable *credulously*: when the argument appears in some preferred extension.

Further discussion and an example making use of each of the above argument status' can be found in [26]. Furthermore, in [57] Dunne and Bench-Capon discuss the computational complexity of a number of decision problems, with respect to classic argumentation frameworks using a single fixed audience, and also in the context of multiple audiences in the value-based framework. However, I conclude the discussion of argumentation frameworks and VAFs here, as the account given above suffices for the purposes required in this thesis.

Bibliography

[1] C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] L. Amgoud and C. Cayrol. On the acceptability of arguments in preference-based argumentation. In G. F. Cooper and S. Moral, editors, *Proccedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI 1998)*, pages 1–7. Morgan-Kaufmann, 1998.

[3] L. Amgoud, N. Maudet, and S. Parsons. Arguments, dialogue, and negotiation. In W. Horn, editor, *Proccedings of the Fourteenth European Conference on Aritificial Intelligence (ECAI 2000)*, pages 338–342, Berlin, Germany, 2000. IOS Press.

[4] Argumentation Service Platform with Integrated Components (ASPIC Project). *Deliverable D1.1 - Review on Argumentation Technology: State of the Art, technical and user requirements*, 2004. Available from the website *http://www.argumentation.org/download.htm*. Accessed on 4th Aug 2005.

[5] J. Argyrakis, S. Gritzalis, and C. Kioulafas. Privacy enhancing technologies: A review. In R. Traunmüller, editor, *Electronic Government (EGOV 2003)*, Lecture Notes in Computer Science 2739, pages 282–287. Springer, Berlin, Germany, 2003.

[6] Aristotle. *Topics*. Clarendon Press, Oxford, UK, 1997. Translated by R. Smith.

[7] Aristotle. *The Nicomachean Ethics*. Oxford University Press, Oxford, UK, 1998. Translated by D. Ross, J. R. Ackrill and J. O. Urmson.

[8] K. Atkinson and T. Bench-Capon. Levels of reasoning with legal cases. In P. E. Dunne and T. Bench-Capon, editors, *ICAIL 2005 Workshop on Argumentation in Artificial Intelligence and Law*, IAAIL Workshop Series, pages 1–11, Nijmegen, The Netherlands, 2005. Wolf Legal Publishers.

[9] K. Atkinson, T. Bench-Capon, and P. McBurney. Attacks on a presumptive argument scheme in multi-agent systems: pre-conditions in terms of beliefs and desires. Technical Report ULCS-04-015, Department of Computer Science, University of Liverpool, UK, 2004.

[10] K. Atkinson, T. Bench-Capon, and P. McBurney. A dialogue game protocol for multi-agent argument for proposals over action. In I. Rahwan, P. Moraitis, and C. Reed, editors, *Proceedings of the First International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2004)*, Lecture Notes in Artificial Intelligence 3366, pages 149–161. Springer, Berlin, Germany, 2004. *An extended version of this paper appears in [15]*.

[11] K. Atkinson, T. Bench-Capon, and P. McBurney. Implementation of a dialogue game for persuasion over action. Technical Report ULCS-04-005, Department of Computer Science, University of Liverpool, UK, 2004.

[12] K. Atkinson, T. Bench-Capon, and P. McBurney. Justifying practical reasoning. In F. Grasso, C. Reed, and G. Carenini, editors, *Proceedings of the Fourth International Workshop on Computational Models of Natural Argument (CMNA 2004)*, pages 87–90, Valencia, Spain, 2004.

[13] K. Atkinson, T. Bench-Capon, and P. McBurney. PARMENIDES: Facilitating democratic debate. In R. Traunmüller, editor, *Electronic Government (EGOV 2004)*, Lecture Notes in Computer Science 3183, pages 313–316. Springer, Berlin, Germany, 2004.

[14] K. Atkinson, T. Bench-Capon, and P. McBurney. Arguing about cases as practical reasoning. In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law (ICAIL 2005)*, pages 35–44, New York, NY, USA, 2005. ACM Press.

[15] K. Atkinson, T. Bench-Capon, and P. McBurney. A dialogue game protocol for multi-agent argument for proposals over action. *Autonomous Agents and Multi-Agent Systems*, 11(2):153–171, 2005. Special Issue on Argumentation in Multi-Agent Systems.

[16] K. Atkinson, T. Bench-Capon, and P. McBurney. Generating intentions through argumentation. In F. Dignum, V. Dignum, S. Koenig, S. Kraus, M. Singh, and M. Wooldridge, editors, *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2005)*, pages 1261–1262, New York, USA, 2005. ACM Press.

[17] K. Atkinson, T. Bench-Capon, and P. McBurney. Multi-agent argumentation for eDemocracy. In M. P. Gleizes, G. Kaminka, A. Nowé, S. Ossowski, K. Tuyls,

and K. Verbeeck, editors, *Proceedings of the Third European Workshop on Multi-Agent Systems*, pages 35–46, 2005.

[18] K. Atkinson, T. Bench-Capon, and P. McBurney. Persuasive political argument. In F. Grasso, C. Reed, and R. Kibble, editors, *Proceedings of the Fifth International Workshop on Computational Models of Natural Argument (CMNA 2005)*, pages 44–51, Edinburgh, Scotland, 2005.

[19] K. Atkinson, T. Bench-Capon, and P. McBurney. Computational representation of practical argument. *Synthese*, 2006. *In press*.

[20] K. Atkinson, T. Bench-Capon, and S. Modgil. Value added: Processing information with argumentation. Technical Report ULCS-05-004, Department of Computer Science, University of Liverpool, UK, 2005.

[21] J. L. Austin. *How to do Things with Words*. Oxford University Press, Oxford, UK, 1962.

[22] T. Bench-Capon. Knowledge based systems applied to law: A framework for discussion. *Knowledge Based Systems and Legal Applications*, pages 329–342, 1991.

[23] T. Bench-Capon. *Practical Legal Expert Systems: the Relation Between a Formalisation of Law and Expert Knowledge*, pages 191–201. Ablex, 1991.

[24] T. Bench-Capon. Specification and implementation of Toulmin Dialogue Game. In *Proceedings of the Eleventh Annual Conference on Legal Knowledge and Information Systems (JURIX 1998)*, pages 5–20, Nijmegen, 1998. GNI.

[25] T. Bench-Capon. Agreeing to differ: Modelling persuasive dialogue between parties without a consensus about values. *Informal Logic*, 22(3):231–245, 2003.

[26] T. Bench-Capon. Persuasion in practical argument using value based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–48, 2003.

[27] T. Bench-Capon, K. Atkinson, and A. Chorley. Persuasion and value in legal argument. *Journal of Logic and Computation*, 15:1075–1097, 2005.

[28] T. Bench-Capon, P. H. Leng, and G. Stainford. A computer supported environment for the teaching of legal argument. *Journal of Law and Information Technology*, 3(3), 1998.

[29] T. Bench-Capon and G. Sartor. Theory based explanation of case law domains. In *Proceedings of the Eighth International Conference on Artificial Intelligence and Law (ICAIL 2001)*, pages 12–21, New York, 2001. ACM Press.

[30] T. Bench-Capon and G. Sartor. A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence*, 150, 2003.

[31] T. Bench-Capon and G. Stainford. PLAID - proactive legal assistance. In *Proceedings of the Fifth International Conference on Artificial Intelligence and Law (ICAIL 1995)*, pages 81–88, New York, 1995. ACM Press.

[32] D. H. Berman and C. D. Hafner. Representing teleological structure in case-based legal reasoning: the missing link. In *Proceedings of the Fourth International Conference on Artificial Intelligence and Law (ICAIL 1993)*, pages 50–59, New York, NY, USA, 1993. ACM Press.

[33] R-J. Beun and R. M. van Eijk. A co-operative dialogue game for resolving ontological discrepancies. In F. Dignum, editor, *Advances in Agent Communication*, Lecture Notes in Artificial Intelligence 2922, pages 349–363. Springer, Berlin, Germany, 2004.

[34] J. Bohman and W. Rehg, editors. *Deliberative Democracy: Essays on Reason and Politics*. MIT Press, Cambridge, MA, USA, 1997.

[35] A. Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93(1–2):63–101, 1997.

[36] M. E. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press, Cambridge, MA, USA, 1987.

[37] M. E. Bratman. What is intention? In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 15–32. MIT Press, Cambridge, MA, USA, 1990.

[38] P. Bretier and D. Sadek. A rational agent as the kernel of a cooperative spoken dialogue system: Implementing a logical theory of interaction. In J. P. Müller, M. Wooldridge, and N. R. Jennings, editors, *Intelligent Agents III*, Lecture Notes in Artificial Intelligence 1193, pages 189–204. Springer, Berlin, Germany, 1997.

[39] J. Breuker and N. den Haan. Separating world and regulation knowledge, where is the logic? In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law (ICAIL 1991)*, pages 92–97, New York, NY, USA, 1991. ACM Press.

[40] G. Brewka and T. F. Gordon. How to buy a Porsche: An approach to defeasible decision making. In *Working Notes of the AAAI-1994 Workshop on Computational Dialectics*, pages 28–38, Seattle, Washington, USA, 1994.

[41] B. G. Buchanan and E. H. Shortliffe. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA, USA, 1984.

[42] D. V. Carbogim, D. Robertson, and J. Lee. Argument-based applications to knowledge engineering. *The Knowledge Engineering Review*, 15(2):119–149, 2000.

[43] C. Castelfranchi. Commitments: From individual intentions to groups and organizations. In V. R. Lesser and L. Gasser, editors, *Proceedings of the First International Conference on Multiagent Systems, San Francisco, California, USA*, pages 41–48. The MIT Press, 1995.

[44] C. Castenfranchi. The theory of social functions. Challenges for multi-agent-based social simlation and multi-agent learning. *Journal of Cognitive Systems Research*, 2(1):5–38, 2001.

[45] C. G. Christie. *The Notion of an Ideal Audience in Legal Argument*. Kluwer Academic Publishers, 2000.

[46] E. Cogan, S. Parsons, and P. McBurney. What kind of argument are we going to have today? In S. Koenig S. Kraus M. P. Singh F. Dignum, V. Dignum and M. Wooldridge, editors, *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2005), Utrecht, The Netherlands.*, pages 544–551, New York, NY, USA, 2005. ACM Press.

[47] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2–3):213–261, 1990.

[48] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 221–255. MIT Press, Cambridge, MA, USA, 1990.

[49] P. R. Cohen and C. R. Perrault. Elements of a plan based theory of speech acts. *Cognitive Science*, 3:177–212, 1979.

[50] J. Coleman. *Risks and Wrongs*. Cambridge University Press, 1992.

[51] F. Dignum, B. Dunin-Kęplicz, and R. Verbrugge. Agent theory for team formation by dialogue. In C. Castelfranchi and Y. Lespérance, editors, *Intelligent Agents VII: Proceedings of the Seventh International Workshop on Agent Theories, Architectures, and Languages (ATAL 2000)*, Lecture Notes in Artificial Intelligence 1986, pages 150–166. Springer, Berlin, Germany, 2000.

[52] F. Dignum, D. Morley, L. Sonenberg, and L. Cavedon. Towards socially sophisticated BDI agents. In *Proceedings of the Fourth International Conference on Multi-agent Systems*, pages 111–118, Boston, USA, 2000.

[53] S. Doutre, T. Bench-Capon, and P. E. Dunne. Explaining preferences with argument positions. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 1560–1561, 2005.

[54] P. M. Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning and logic programming. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI 1993)*, pages 852–859, 1993.

[55] P. M. Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[56] P. E. Dunne and T. Bench-Capon. Coherence in finite argument systems. *Artificial Intelligence*, 141(1):187–203, 2002.

[57] P. E. Dunne and T. Bench-Capon. Complexity in value-based argument systems. In *The European Conference on Logics in Artificial Intelligence (JELIA 2004)*, pages 360–371, 2004.

[58] R. M. van Eijk. *Programming Languages for Agent Communications*. PhD thesis, University of Utrecht, Utrecht, The Netherlands, 2000.

[59] R. M. van Eijk, F. S. de Boer, W. van der Hoek, and J-J. Ch. Meyer. Operational semantics for agent communication languages. In F. Dignum and M. Greaves, editors, *Issues in Agent Communication*, Lecture Notes in Artificial Intelligence 1916, pages 80–95. Springer, Berlin, Germany, 2000.

[60] S. Fatima, M. Wooldridge, and N. Jennings. An agenda based framework for multi-issues negotiation. *Artificial Intelligence*, 152(1):1–45, 2004.

[61] FIPA. Communicative Act Library Specification. Technical Report XC00037H, Foundation for Intelligent Physical Agents, 10th August 2001. Available from the website *http://www.fipa.org*. Accessed on 4th August 2005.

[62] J. Fox. Agents in healthcare. In J. L. Nealon and A. Moreno, editors, *Applications of Software Agent Technology in the Health Care Domain*, Whitestein Series in Software Agent Technologies. Birkhuser Verlag, Basel, 2003.

[63] J. Fox and D. W. Glasspool. Knowledge and argument in clinical decision-making. *Multidisciplinary Approaches to Theory in Medicine*, Studies in Multidisciplinarity, 2005. *In press*.

[64] J. Fox and S. Parsons. On using arguments for reasoning about actions and values. In *Proceedings of the AAAI Spring Symposium on Qualitative Preferences in Deliberation and Practical Reasoning*, Stanford, USA, 1997.

[65] J. Fox and S. Parsons. Arguing about beliefs and actions. In A. Hunter and S. Parsons, editors, *Applications of Uncertainty Formalisms*, pages 266–302, Berlin, Germany, 1998. Springer.

[66] H. G. Frankfurt. Freedom of the will and the concept of a person. *Journal of Philosophy*, 68(1):5–20, 1971.

[67] G. Frege. *Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reinen Denkens.* Halle, 1879.

[68] M. P. Georgeff and A. L. Lansky. Reactive reasoning and planning. In *Proceedings of the Sixth International Conference on Artificial Intelligence (AAAI 1987)*, pages 677–682, Seattle, WA, 1987.

[69] R. Goldblatt. *Topoi: The Categorial Analysis of Logic*. North-Holland, Amsterdam, The Netherlands, 1979.

[70] T. F. Gordon. The Pleadings Game: An exercise in computational dialectics. *Artificial Intelligence and Law*, 2:239–292, 1994.

[71] T. F. Gordon and N. I. Karacapilidis. The Zeno argumentation framework. In *Proceedings of Sixth International Conference on Artificial Intelligence and Law (ICAIL 2003)*, pages 10–18, New York, 1997. ACM Press.

[72] T. F. Gordon and G. Richter. Discourse support systems for deliberative democracy. In R. Traunmüller and K. Lenk, editors, *Electronic Government (EGOV 2002)*, Lecture Notes in Computer Science 2456, pages 238–255. Springer, Berlin, Germany, 2002.

[73] F. Grasso, A. Cawsey, and R. Jones. Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *International Journal of Human-Computer Studies*, 53:1077–1115, 2000.

[74] K. Greenwood, T. Bench-Capon, and P. McBurney. Structuring dialogue between the People and their representatives. In R. Traunmüller, editor, *Electronic Government (EGOV 2003)*, Lecture Notes in Computer Science 2739, pages 55–62. Springer, Berlin, Germany, 2003.

[75] K. Greenwood, T. Bench-Capon, and P. McBurney. Towards a computational account of persuasion in law. In *Proceedings of the Ninth International Conference on Artificial Intelligence and Law (ICAIL 2003)*, pages 22–31, New York, NY, USA, 2003. ACM Press.

[76] W. Grennan, editor. *Informal Logic. Issues and Techniques*. McGill Queens University Press, Canada, 1997.

[77] C. A. Gunter. *Semantics of Programming Languages: Structures and Techniques*. Foundations of Computing Series. MIT Press, Cambridge, MA, USA, 1992.

[78] J. Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, Cambridge, MA, USA, 1996. (Translation by W. Rehg).

[79] J. Hage. *Reasoning with Rules*. Kluwer Academic Publishers: Dordrecht, 1997.

[80] C. L. Hamblin. *Fallacies*. Methuen, London, UK, 1970.

[81] D. Hitchcock. Fallacies and formal logic in Aristotle. *History and Philosophy of Logic*, 21:207–221, 2000.

[82] D. Hitchcock. Pollock on practical reasoning. *Informal Logic*, 22(3):247–256, 2002.

[83] D. Hume. *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects*. 1739-40. Reprinted in 2000 by Oxford University Press, UK.

[84] A. Hunter. Making argumentation more believable. In *Proceedings of the Ninteenth American National Conference on Artificial Intelligence (AAAI 2004)*, pages 269–274, Cambridge, MA, USA, 2004. MIT Press.

[85] A. Hunter. Towards higher impact argumentation. In *Proceedings of the Ninteenth American National Conference on Artificial Intelligence (AAAI 2004)*, pages 275–280, Cambridge, MA, USA, 2004. MIT Press.

[86] C. Hurt, J. Fox, J. Bury, and V. Saha. Computerised advice on drug dosage decisions in childhood leukaemia: a method and a safety strategy. In M. Dojat, E. Keravnou, and P. Barahona., editors, *Artificial Intelligence in Medicine: Ninth Conference on Artificial Intelligence, in Medicine in Europe, AIME 2003*, Lecture Notes in Artificial Intelligence 2780, pages 158–162. Springer, Berlin, Germany, 2003.

[87] D. Ince. *Developing Distributed and E-Commerce Applications*. Addison Wesley, 2002.

[88] M. Jarke, M. T. Jelassi, and M. F. Shakun. MEDIATOR: Towards a negotiation support system. *European Journal of Operational Research*, 31:314–333, 1987.

[89] H. Kamp and U. Reyle. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic, Dordrecht, The Netherlands, 1993. Two Volumes.

[90] I. Kant. *Groundwork of the Metaphysics of Morals*. Harper Torchbooks, New York, USA, 1785. Reprinted in 1964.

[91] J. Katzav and C. Reed. On argumentation schemes and the natural classification of arguments. *Argumentation*, 18(2):239–259, 2004.

[92] A. J. P. Kenny. *Practical Reasoning and Rational Appetite*. 1975. Reprinted in [138].

[93] M. Kienpointner. Towards a typology of argument schemes. In *Proceedings of ISSA 1986*, Amsterdam, The Netherlands, 1986. Amsterdam University Press.

[94] R. Kowalczyk and V. Bui. On constraint-based reasoning in e-negotiation agents. In F. Dignum and U. Cortés, editors, *Agent-Mediated Electronic Commerce III*, Lecture Notes in Computer Science 2003, pages 31–46. Springer, Berlin, Germany, 2001.

[95] R. Kowalski and F. Toni. Abstract argumentation. *Artificial Intelligence and Law, Special Issue on Logical Models of Argumentation*, 1996.

[96] S. Kraus. *Strategic Negotiation in Multi-Agent Environments*. MIT Press, Cambridge, MA, USA, 2001.

[97] P. Krause, S. Ambler, M. Elvang-Gorensson, and J. Fox. A logic for argumentation for reasoning under uncertainty. *Computational Intelligence*, 11(1):113–131, 1995.

[98] Y. Labrou, T. Finin, and Y. Peng. Agent communication languages: The current landscape. *IEEE Intelligent Systems*, 14(2):45–52, 1999.

[99] L. Lindahl. Deduction and justification in the law. The role of legal terms and concepts. *Ratio Juris*, 17(2):182–202, 2004.

[100] A. R. Lodder. *DiaLaw: On Legal Justification and Dialog Games*. PhD thesis, University of Maastricht, Maastricht, The Netherlands, 1998.

[101] R. Lührs, S. Albrecht, M. Lübcke, and B. Hohberg. How to grow? online consultation about growth in the city of Hamburg: methods, techniques, success factors. In R. Traunmüller, editor, *Electronic Government (EGOV 2003)*, Lecture Notes in Computer Science 2739, pages 79–84. Springer, Berlin, Germany, 2003.

[102] J. D. MacKenzie. Question-begging in non-cumulative systems. *Journal of Philosophical Logic*, 8:117–133, 1979.

[103] C. C. Marshall. Representing the structure of a legal argument. In *Proceedings of the Second International Conference on Artificial Intelligence and Law (ICAIL 1989)*, pages 121–127, 1989.

[104] N. Maudet and B. Chaib-Draa. Commitment-based and dialogue-game-based protocols. *The Knowledge Engineering Review*, 17(2):157–179, 2002.

[105] P. McBurney and S. Parsons. Intelligent systems to support deliberative democracy in environmental regulation. *ICT Law*, 10(1):33–43, 2001.

[106] P. McBurney and S. Parsons. Representing epistemic uncertainty by means of dialectical argumentation. *Annals of Mathematics and Artificial Intelligence*, 32(1–4):125–169, 2001.

[107] P. McBurney and S. Parsons. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of Logic, Language and Information*, 11(3):315–334, 2002.

[108] P. McBurney and S. Parsons. A denotational semantics for deliberation dialogues. In N. R. Jennings, C. Sierra, L. Sonenberg, and M. Tambe, editors, *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2004), New York City, NY, USA*, pages 86–93, New York City, NY, USA, 2004. ACM Press.

[109] P. McBurney, S. Parsons, and M. Wooldridge. Desiderata for agent argumentation protocols. In C. Castelfranchi and W. L. Johnson, editors, *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002)*, pages 402–409, Bologna, Italy, 2002. ACM Press: New York, USA.

[110] P. McBurney, R. M. van Eijk, S. Parsons, and L. Amgoud. A dialogue-game protocol for agent purchase negotiations. *Autonomous Agents and Multi-Agent Systems*, 7(3):235–273, 2003.

[111] L. T. McCarty. An implementation of Eisner v. Macomber. In *Proceedings of the Fifth International Conference on Artificial Intelligence and Law (ICAIL 1995)*, pages 276–286, New York, NY, USA, 1995. ACM Press.

[112] L. T. McCarty and M. S. Sridharan. A computational theory of legal argument. Technical Report LRP-TR-13, Computer Science Department, Rutgers University, 1982.

[113] J-J. Ch. Meyer and W. van der Hoek. *Epistemic Logic for Computer Science and Artificial Intelligence*. Cambridge Tracts in Theoretical Computer Science 41. Cambridge University Press, Cambridge, UK, 1995.

[114] F. I. Michelman. Conceptions of democracy in American Constitutional argument: the case of pornography regulation. *Tennessee Law Review*, 56:291–319, 1989.

[115] E. Millgram, editor. *Varieties of Practical Reasoning*. MIT Press: A Bradford Book, Cambridge, MA, USA, 2001.

[116] D. J. Moore. *Dialogue Game Theory for Intelligent Tutoring Systems*. PhD thesis, Leeds Metropolitan University, Leeds, UK, 1993.

[117] A. Omicini, A. Ricci, and M. Viroli. Agens faber: Toward a theory of artifacts for mas. In *First International Workshop on Coordination and Organisation (CoOrg 2005), Namur, Belgium*, Electronic Notes in Theoretical Computer Science. 2005. *In press.*

[118] R. Parikh. Social software. *Synthese*, 132:187–211, 2002.

[119] S. Parsons. *Qualitative Methods for Reasoning Under Uncertainty*. MIT Press, Cambridge, MA, USA, 2001.

[120] S. Parsons, C. Sierra, and N. R. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261–292, 1998.

[121] R. Patil, R. F. Fikes, P. F. Patel-Schneider, D. McKay, T. Finin, T. Gruber, and R. Neches. The DARPA knowledge sharing effort: Progress report. In B. Nebel, C. Rich, and W. Swartout, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference*, pages 777–788. Morgan Kaufmann, CA, USA, 1992.

[122] C. Perelman. *Justice, Law, and Argument*. D. Reidel Publishing Company, Dordrecht, Holland, 1980.

[123] C. Perelman and L. Olbrechts-Tyteca. *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press, Notre Dame, IN, USA, 1969.

[124] J. Pitt and A. Mamdani. Some remarks on the semantics of FIPA's Agent Communications Language. *Autonomous Agents and Multi-Agent Systems*, 2(4):333–356, 1999.

[125] M. E. Pollack. The uses of plans. *Artificial Intelligence*, 57(1):43–68, 1992.

[126] J. Pollock. *Cognitive Carpentry: A Blueprint for How to Build a Person*. MIT Press, MA, USA, 1995.

[127] J. Pollock. Rational cognition in OSCAR. In N. Jennings and Y. Lesperance, editors, *Intelligent Agents VI. Agent Theories, Architectures, and Languages*, Lecture Notes in Artificial Intelligence 1757, pages 49–58. Springer, Berlin, Germany, 1999.

[128] H. Prakken. An exercise in formalising teleological reasoning. In *Proceedings of the Thireteenth Annual Conference on Legal Knowledge and Information Systems (JURIX 2000)*, pages 49–58, Amsterdam, The Netherlands, 2000. IOS Press.

[129] H. Prakken. A study of accrual of arguments, with applications to evidential reasoning. In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law (ICAIL 2005)*, pages 85–94, New York, USA, 2005. ACM Press.

[130] H. Prakken, C. Reed, and D. N. Walton. Argumentation schemes and generalisations in reasoning about evidence. In *Proceedings of the Ninth International Conference on Artificial Intelligence and Law (ICAIL 2003)*, pages 32–41, New York, USA, 2003. ACM Press.

[131] H. Prakken, C. Reed, and D. N. Walton. Argumentation schemes and burden of proof. In F. Grasso, C. Reed, and G. Carenini, editors, *Proceedings of the Fourth International Workshop on Computational Models of Natural Argument (CMNA 2004)*, pages 81–86, Valencia, Spain, 2004.

[132] A. Prosser, R. Kofler, and R. Krimmer. Deploying electronic democracy for public corporations. In R. Traunmüller, editor, *Electronic Government (EGOV 2003)*, Lecture Notes in Computer Science 2739, pages 234–239. Springer, Berlin, Germany, 2003.

[133] I. Rahwan. *Interest-based Negotiation in Multi-Agent Systems*. PhD thesis, University of Melbourne, Melbourne, Australia, 2004.

[134] I. Rahwan, S. Ramchurn, N. Jennings, P. McBurney, S. Parsons, and L. Sonenberg. Argumentation-based negotiation. *The Knowledge Engineering Review*, 18(4):343–375, 2004.

[135] I. Rahwan, L. Sonenberg, and F. Dignum. Towards interest-based negotiation. In J. Rosenschein, T. Sandholm, M. Wooldridge, and M. Yokoo, editors, *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2003)*, pages 773–780. ACM Press, 2003.

[136] H. Raiffa. *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. Addison-Wesley, Reading, MA, 1970.

[137] S. Ramchurn, N. Jennings, and C. Sierra. Persuasive negotiation for autonomous agents: a rhetorical approach. In C. Reed, F. Grasso, and G. Carenini, editors, *Proceedings of the Fifth International Workshop on Computational Models of Natural Argument (CMNA 2003)*, pages 9–17. AAAI Press, 2003.

[138] J. Raz, editor. *Practical Reasoning*. Oxford University Press, Oxford, UK, 1978.

[139] A. Rector. Medical informatics. In D. McGuinness D. Nardi F. Baader, D. Calvanese and P.F. Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[140] C. Reed and G. Rowe. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 14(3–4):961–980, 2004.

[141] C. Reed and D. Walton. Towards a formal and implemented model of argumentation schemes in agent communication. *Autonomous Agents and Multi-Agent Systems*, 11(2):173–188, 2005. Special Issue on Argumentation in Multi-Agent Systems.

[142] H. S. Richardson. *Practical Reasoning about Final Ends*. Cambridge University Press, Cambridge, UK, 1994.

[143] E. L. Rissland and K. D. Ashley. A note on dimensions and factors. *AI and Law*, 10(1–3):65–77, 2002.

[144] H. W. J. Rittel and M. M. Webber. Dilemas in a general theory of planning. *Policy Sciences*, pages 155–169, 1973.

[145] J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, UK, 1969.

[146] J. R. Searle. *Rationality in Action*. MIT Press, Cambridge, MA, USA, 2001.

[147] M. J. Sergot, F. Sadri, R. A. Kowalski, F. Kriwaczek, P. Hammond, and H. T. Cory. The British Nationality Act as a logic program. *Communications of the ACM*, 29(5):370–386, 1986.

[148] G. L. S. Shackle. *Decision, Order and Time in Human Affairs*. Cambridge University Press, Cambridge, UK, 1961. 2nd edition 1969.

[149] G. Shafer. *Mathematical Theory of Evidence*. Princeton University Press, USA, 1976.

[150] C. Sierra, N. R. Jennings, P. Noriega, and S. Parsons. A framework for argument based negotiation. In A. Rao M. Singh and M. Wooldridge, editors, *Intelligent Agents IV*, Lecture Notes in Artificial Intelligence 1365, pages 177–192. Springer, Berlin, Germany, 1998.

[151] M. P. Singh. A social semantics for agent communication languages. In *Issues in Agent Communication*, Lecture Notes In Computer Science 1916, pages 31–45. Springer, Berlin, Germany, 2000.

[152] E. Smith and A. Macintosh. E-Voting: powerful symbol of e-democracy. In R. Traunmüller, editor, *Electronic Government (EGOV 2003)*, Lecture Notes in Computer Science 2739, pages 240–245. Springer, Berlin, Germany, 2003.

[153] K. Sycara. Persuasive argumentation in negotiation. *Theory and Decision*, 28(3):203–242, 1990.

[154] V. Tamma and T. Bench-Capon. A conceptual model to facilitate knowledge sharing in multi-agent systems. In *Proceedings of Autonomous Agents 2001 Workshop on Ontologies in Agent Systems (OAS 2001)*, pages 69–76, Montreal, Canada, 2001.

[155] V. Tamma, I. Blacoe, B. Lithgow-Smith, and M. Wooldridge. SERSE: Searching for semantic web content. In R. López de Mántaras and L. Saitta, editors, *Proceedings of the Sixteenth European Conference on Artificial Intelligence (ECAI 2004)*, pages 63–67, 2004.

[156] R. D. Tennent. *Semantics of Programming Languages*. Prentice-Hall, Hemel Hempstead, UK, 1991.

[157] S. Toulmin. *The Uses of Argument*. Cambridge University Press, Cambridge, UK, 1958.

[158] A. Trollope. *The American Senator*. Oxford University Press, Oxford, UK, 1986.

[159] S. W. Tu and M. A. Musen. Representation formalisms and computational methods for modeling guideline-based patient care. In M. Mussen M. Stefanelli B. Heller, M. Loffler, editor, *Proceedings of First European Workshop on Computer-based Support for Clinical Guidelines and Protocols*, pages 125–142, Leipzig, Germany, 2000. IOS Press.

[160] A. Valente. *Legal Knowledge Engineering: A modelling approach*. IOS Press, Amsterdam, 1995.

[161] B. Verheij. Dialectical argumentation with argumentation schemes: An approach to legal logic. *Artificial Intelligence and Law*, 11:167–195, 2003.

[162] G. A. W. Vreeswijk and H. Prakken. Credulous and sceptical argument games for preferred semantics. In G. Brewka M. Ojeda-Aciego, Inma P. de Guzmán and L. Moniz Pereira, editors, *Logics in Artificial Intelligence: Proceedings of the Seventh European Workshop*, Lecture Notes in Artificial Intelligence 1919, pages 239–253. Springer, Berlin, Germany, 2000.

[163] D. N. Walton. *Practical Reasoning: Goal-Driven, Knowledge-Based, Action-Guiding Argumentation*. Rowman and Littlefield, Savage, Maryland, USA, 1990.

[164] D. N. Walton. *Argument Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1996.

[165] D. N. Walton. *Appeal to Expert Opinion*. Penn State Press, University Press, USA, 1997.

[166] D. N. Walton. *The New Dialectic: Conversational Contexts of Argument*. University of Toronto Press, Toronto, Ontario, Canada, 1998.

[167] D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Press, Albany, NY, USA, 1995.

[168] M. Wooldridge. *Reasoning about Rational Agents*. MIT Press, Cambridge, MA, USA, 2000.

[169] M. Wooldridge. *An Introduction to MultiAgent Systems*. John Wiley and Sons, New York, NY, USA, 2001.

[170] M. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

[171] R. R. Yager, M. Fedrizzi, and J. Kacprzyk, editors. *Advances in the Dempster-Shafer Theory of Evidence*. John Wiley and Sons, 1994.

[172] T. Yuan. *Human Computer Debate, a Computational Dialectics Approach*. PhD thesis, Leeds Metropolitan University, Leeds, UK, 2004.

[173] J. Zeleznikow and A. Stranieri. The split-up system: integrating neural networks and rule-based reasoning in the legal domain. In *Proceedings of the Fifth International Conference on Artificial Intelligence and Law (ICAIL 1995)*, pages 185–194, New York, NY, USA, 1995. ACM Press.